Sa

NAVAL RESEARCH LOGISTICS QUARTERLY

POOL

OFFICE OF NAVAL RESEARCH

CONTENTS

ARTICLES

Risk and Hazard in Logistics Planning — 217
Colonel Harry A. Sachaklian

Comments on the Morgenstern Model — 225
Lieutenant Commander J. C. Busby

On the Meaning and Use of a Capacity Concept — 237
Edward Zabel

Issue Priority: Last in First Out (LIFO) vs. First in First Out (FIFO) as a Method of Issuing Items from Supply Storage — 251
J. A. Green

Cycling in the Dual Simplex Algorithm — 269
E. M. L. Beale

Computation of Maximal Flows in Networks — 277
D. R. Fulkerson and G. B. Dantzig

The Structure of Dynamic Programming Models — 285
S. Karlin

NEWS AND MEMORANDA — 295

RECENT PUBLICATIONS — 299

DEPOSITORY
APR 30 1956
Los Angeles Public Library

VOL. 2, NO. 4 DECEMBER 1955

NAVEXOS P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

Rear Admiral H. E. Eccles, USN (Retired)
The George Washington University

O. Morgenstern
Princeton University

Commander G. S. Essenwine, USN
Office of Naval Research

F. D. Rigby
Office of Naval Research

Alan J. Hoffman, *Managing Editor*
Office of Naval Research
Washington 25, D. C.

ASSOCIATE EDITORS

D. Blackwell, *University of California*
R. M. Bowstrom, *Captain, SC, USN*
E. W. Cannon, *National Bureau of Standards*
G. B. Dantzig, *The Rand Corporation*
T. B. Evans, *Colonel, USA*
M. M. Flood, *Columbia University*
J. D. Hayes, *Rear Admiral, USN (Retired)*
E. B. Henry, Jr., *Commander, USN*
J. Laderman, *Office of Naval Research Br. Office*
A. T. Magnell, *Captain, SC, USN*
J. Marschak, *Cowles Foundation*
T. C. Roberts, *Commander, SC, USN*
M. E. Rose, *Office of Naval Research*
H. P. Sachaklian, *Colonel, USAF*
E. K. Scofield, *Commander, SC, USN*
J. S. Skoczylas, *Lt. Colonel, USMC*
E. D. Stanley, *Captain, SC, USN*
H. Strock, *Commander, SC, USN*
R. M. Thrall, *University of Michigan*
C. B. Tompkins, *University of California*
A. W. Tucker, *Princeton University*
R. E. Williams, *Commander, SC, USN*
M. A. Woodbury, *The George Washington University*

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Manuscripts and other items for publication should be sent to the Managing Editor at the above address. Authors will receive 50 free reprints of their published papers.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the month of March, June, September, and December and can be purchased from the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. Subscription Price: $1.50 a year in the U. S. and Canada, $2.00 elsewhere; 50¢ for a single copy. Reprints of an individual article can be purchased in quantities of 100 or more. Requests for the purchase price of reprints of a particular article should be sent to the Superintendent of Documents.

The views and opinions expressed in the articles in this Quarterly are those of the authors and not necessarily those of the Office of Naval Research.

The printing of this publication was approved by the Director of the Bureau of the Budget, July 26, 19

NAVAL RESEARCH LOGISTICS QUARTERLY

OFFICE OF NAVAL RESEARCH

VOLUME 2

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

Rear Admiral H. E. Eccles, USN (Retired)
The George Washington University

O. Morgenstern
Princeton University

Commander G. S. Essenwine, USN
Office of Naval Research

F. D. Rigby
Office of Naval Research

Alan J. Hoffman, *Managing Editor*
Office of Naval Research
Washington 25, D. C.

ASSOCIATE EDITORS

D. Blackwell, *University of California*
R. M. Bowstrom, *Captain, SC, USN*
E. W. Cannon, *National Bureau of Standards*
G. B. Dantzig, *The Rand Corporation*
T. B. Evans, *Colonel, USA*
M. M. Flood, *Columbia University*
J. D. Hayes, *Rear Admiral, USN (Retired)*
E. B. Henry, Jr., *Commander, USN*
J. Laderman, *Office of Naval Research Br. Office*
A. T. Magnell, *Captain, SC, USN*
J. Marschak, *Cowles Foundation*
T. C. Roberts, *Commander, SC, USN*
M. E. Rose, *Office of Naval Research*
H. P. Sachaklian, *Colonel, USAF*
E. K. Scofield, *Commander, SC, USN*
J. S. Skoczylas, *Lt. Colonel, USMC*
E. D. Stanley, *Captain, SC, USN*
H. Strock, *Commander, SC, USN*
R. M. Thrall, *University of Michigan*
C. B. Tompkins, *University of California*
A. W. Tucker, *Princeton University*
R. E. Williams, *Commander, SC, USN*
M. A. Woodbury, *The George Washington University*

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Manuscripts and other items for publication should be sent to the Managing Editor at the above address. Authors will receive 50 free reprints of their published papers.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. Subscription Price: $1.50 a year in the U. S. and Canada, $2.00 elsewhere; 50¢ for a single copy. Reprints of an individual article can be purchased in quantities of 100 or more. Requests for the purchase price of reprints of a particular article should be sent to the Superintendent of Documents.

The views and opinions expressed in the articles in this Quarterly are those of the authors and not necessarily those of the Office of Naval Research.

The printing of this publication was approved by the Director of the Bureau of the Budget, July 26, 1955

CONTENTS OF VOLUME II

ARTICLES	Pages
ANTOSIEWICZ, H. A. Analytic Study of War Games	181-208
BEALE, E. M. L. Cycling in the Dual Simplex Algorithm	269-275
BRYAN, J. G., WADSWORTH, G. P., and WHITIN, T. M. A Multi-Stage Inventory Model	25-37
BUSBY, J. C., Commander. Comments on the Morgenstern Model	225-235
CURTIS, Captain Ira N. Logistics Without Storage	125-128
DANSKIN, J. M. Mathematical Treatment of a Stockpiling Problem	99-109
DANTZIG, G. B. and FULKERSON, D. R. Computation of Maximal Flows in Networks	277-283
DAVIS, Harold. A Mathematical Evaluation of a Work Sampling Technique	111-117
ECCLES, Rear Admiral Henry E. (Retired). Some Command Problems and Decisions	9-15
FULKERSON, D. R. and DANTZIG, G. B. Computation of Maximal Flows in Networks	277-283
GASS, Saul and SAATY, Thomas. The Computational Algorithm for the Parametric Objective Function	39-45
GEISLER, M. A. and PETERSEN, J. W. The Costs of Alternative Air Base Stocking and Requisitioning Policies	69-81
GREENWOOD, J. A. Issue Priority: Last in First Out (LIFO) vs. First in First Out (FIFO) as a Method of Issuing Items from Supply Storage	251-268
ISAACS, Rufus. The Problem of Aiming and Evasion	47-67
KARLIN, Samuel. The Structure of Dynamic Programming Models	285-294
KELLEY, J. E., Jr. A Dynamic Transportation Model	175-180
KUHN, H. W. The Hungarian Method for the Assignment Problem	83-97
MACKENZIE, Captain D. C. Contractor Performance	17-23
MCSHANE, Vice Admiral R. E. (Retired). Science and Logistics	1-7
MILLS, H. D. Organized Decision Making	137-143
MORGENSTERN, Oskar. Note on the Formulation of the Theory of Logistics	129-136
PETERSEN, J. W., and GEISLER, M. W. The Costs of Alternative Air Base Stocking and Requisitioning Policies	69-81
SAATY, Thomas and GASS, Saul. The Computational Algorithm for the Parametric Objective Function	39-45
SACHAKLIAN, Colonel H. A. Risk and Hazard in Logistics Planning	217-224
SIMPSON, J. R. A Formula for Decisions on Retention or Disposal of Excess Stock	145-155
WADSWORTH, G. P., BRYAN, J. G., and WHITIN, T. M. A Multi-Stage Inventory Model	25-37

ARTICLES Pages

WHITIN, T. M., BRYAN, J. G., and WADSWORTH, G. P. A Multi-Stage Inventory Model 25-37

WHITIN, T. M., and YOUNGS, J. W. T. A Method for Calculating Optimal Inventory Levels and Delivery Time 157-173

YOUNGS, J. W. T. and WHITIN, T. M. A Method for Calculating Optimal Inventory Levels and Delivery Time 157-173

ZABEL, Edward. On the Meaning and Use of a Capacity Concept 237-250

MISCELLANEOUS

News and Memoranda 119-121, 209-211, 295-297
123-124, 213-215, 299-300

RISK AND HAZARD IN LOGISTICS PLANNING

Colonel Harry A. Sachaklian
U. S. Air Force

This paper discusses the significance of the concepts of risk and hazard in areas of military decision.

The customary preface to any treatise on military science, particularly as concerns its planning aspects, is to comment on its inexact nature. This has been pointed out so often as to approach being platitudinous. It becomes necessary, then, to make clear that reference to the inexactness of military science herein is an accusation and not an apology.

Whereas inexactness, in truth, is a condition of military planning, its acceptance as an irreducible condition serves only the interests of incompetence since specific responsibility for error can thereby be avoided.

The growing concern in the nation as a whole about the cost of the military establishment is resulting in increasingly sharp criticism of errors in military planning. This criticism has reached the stage where formal proposals are being made to remove certain military functions from military hands and place them in civilian agencies to be created for this purpose. These proposals are being made in the face of clear-cut evidence that military agencies are equally concerned with these problems and are making headway in their resolution.

Whereas such proposals add nothing to accuracy and merely change the error-makers from military to civilian, they are strong evidence of the widespread dissatisfaction with present procedure. The requirement for a review and a strengthening of planning concepts and procedures is urgent, not only because of the unacceptable inexactness from a military point of view, but because the understandable reaction to continuing error, especially in its materiel requirements manifestations, may result in organization changes that by their very nature would increase, not decrease, the error.

There is no single area in military science where error is so readily discoverable as that portion of military science concerned with determination of materiel requirements. The oddity here is that the real evidence of inherent error is the inability of the determiner to prove the accuracy of his calculations and not the ability of the critic to occasionally isolate specific examples of inaccuracy.

The determiner is therefore wrong because the only evidence he can now offer to substantiate his calculations is that the resources procured as a result of his calculations were consumed. Since armed forces are monsters with insatiable appetities that can and will consume all they can get, evidence of consumption is hardly proof of need.

Nor can evidence of nonuse be universally accepted as proof of no need. The very concept of "forces-in-being" is based on the belief that the existence and capability of these forces will preclude the necessity for their use.

Assuming that the basic purpose of logistics support is to contribute to operational effectiveness it necessarily follows that proof of need must be measured in terms of such effectiveness. Armed forces can absorb a tremendous amount of logistics support without any appreciable increase in operational effectiveness. In fact, there is a point beyond which the

volume of logistics support actually decreases operational effectiveness since the physical weight of such excess support directly reduces the mobility and flexibility of the armed forces in addition to depriving the operational element of the resources otherwise available to it in order to manage the impedimenta.

Any review of planning concepts with an object of strengthening materiel requirements determination procedures must begin by exploring the relationship of logistics to strategy in order to find a means of measuring the contribution of logistics support to operational effectiveness.

Operational effectiveness is a compound of strategy, state of training, and logistics. Strategy is a scheme of maneuver expressed as a series of decisions concerning the employment and deployment of armed forces. State of training is the level of the collective skill of the armed forces in carrying out these decisions. Logistics is the provision of the resources required by the strategy as conditioned by the state of training.

The whole scope of military decisions can be reduced to eight separate and distinct areas, all related to each other and all consisting of a multitude of subordinate decisions.

The entire purpose of an armed force created for defense purposes is predicated upon the existence of a military threat to the security of the nation. The size and nature of this defensive armed force is directly conditioned by the size and nature of the threat. The balancing, neutralizing, or reduction of this threat becomes the prime task of the defensive armed force and consequently the material manifestations of the threat can be called the TARGET.

The first and most important series of decisions to make concerns this target, such as, what is it, where is it, how well armed is it and most important of all, what is to be done to it. The decision as to what is to be done to the target in effect guides and limits all the rest of the military decisions.

The second decision concerns the WEAPON. Having decided what needs to be done to the target, the subsequent question is a qualitative and quantitative analysis of the weapons required to do to the target what needs to be done.

The third decision concerns the FORCE required to apply the selected weapon or series of weapons to the target. This series of decisions is the most complex and involves such questions as deployment, scheme of employment and training required to perform the mission of applying weapons to targets.

The fourth decision concerns the BASE or bases required to house and launch the force that will carry the weapon to the target. The subject of bases can be as simple in nature as a foxhole or as vast and complex as a nation in toto.

These four areas of military decision; target, weapon, force, base, are clearly and unequivocally elements of strategic decision in that the factors upon which the decisions are based are derived directly from the size and nature of the material manifestations of the threat to the security of the nation.

The determination of the most desirable targets, weapons, forces and bases is tne fundamental function of the operations planner. The elements of state of training and logistics enter into these calculations after they are made and then only as a limiting factor. They do not change the essentiality of these elements of strategic decision. They only influence the degree to which these elements are attainable.

The four additional areas of military decision that must be made to accomplish the action required in the elements of strategic decision are the elements of logistics decision which are in most part predicated upon and derived from the elements of strategic decision.

The first of these is DISTRIBUTION. This is defined as the systematic deployment of resources to meet the strategy implicit in the nature and location of the forces and bases.

The second is TRANSPORTATION. Transportation in this sense is considered to be an integrated system through which distribution can be accomplished in the volume and time required by the strategy implicit in the nature and location of forces and bases.

The third is PROCUREMENT. This is also a system and consists of a procedure by which title to resources passes from one agency to another.

The fourth is PRODUCTION. Production again is a system. It consists of transforming material things into items required by the strategy implicit in the nature and location of forces and bases.

In short, the elements of logistics decision concern the establishment of systems through which the material requirements for defense are produced, procured, transported, and distributed to bases to enable a force to apply weapons to targets.

Logistics systems and strategic systems (the target, weapon, force, base complex) are both composed of personnel and materiel, cost money to operate and therefore compete for resources. Obviously the reduction of the cost of logistics systems permits the greater allocation of resources to the strategic systems within any given economic limitation.

Certainly the foregoing hypothetical expression of the processes of military decision is an oversimplification. Yet its value is in its explanation of relationships among these functions rather than as an exposition of the separate functions themselves. Its utility as a basis for requirements determination is in direct proportion to the acceptability of this concept of relationship.

The relation of logistics to strategy is closest in the field of requirements determination. A proposed strategy, when translated into specific requirements, can thereby be tested for logistics feasibility and if necessary, adjusted accordingly. Logistics, then, not only is derived from strategy but by virtue of its limiting effect contributes to the determination of that from which it is derived. Therefore the treatment of logistics as the natural son of strategy is at least a biological contradiction since no natural son has ever been privileged to assist in the selection of his father.

Military logistics requirements, by historic usage, pertain generally to four different areas; supply, maintenance, construction, and transportation. Such fields as communications in themselves are only separate logistics problems, in that they also require supply, maintenance, construction, and transportation.

It would seem to be most useful to classify logistics requirements for these four areas in terms of the use to which they are being put. If this were done, there would be three broad classifications of logistics requirements:

a. Requirements of and for personnel.
b. Requirements of and for equipment.
c. Requirements of and for bases or installations.

Each of these three can in turn be further subdivided into requirements that are utilized at a relatively even rate and those utilized at a relatively erratic rate.

This basic classification can readily be converted into requirements determination procedure in that it indicates the types of requirements capable of being reduced to reasonably accurate factors and those that must be determined by other means.

Requirements to support personnel are relatively easy to predict since they can be readily factored on a population basis. Such things as food, clothing, and shelter can be

forecast either by usage data or by simple calculation. In usual circumstances these requirements are utilized at a relatively even rate. Certain wartime conditions may cause erratic usage rates but in comparison to the total military requirement, such variations are small enough to handle on a local basis and have little effect on the overall requirement.

Requirements to support equipment may in part be derived from usage data particularly as a guide to replacement of component parts or the equipment itself caused by wear due to normal usage. However, the relatively unpredictable nature of battle damage complicates the problem considerably. Requirements for certain items, such as ammunition, would vary widely not only from peace to war but also within war from area to area.

Requirements to support bases or installations are predictable from usage data or from the specifications of the installation for normal operating requirements. However, such predictions are inapplicable when such bases or installations are subjected to enemy attack.

Thereafter, the entire problem of requirements determination can be considered as having six integrants:

a. WHAT is needed? This is the specification of an item, so designated as to be readily identifiable.
b. WHERE is it needed? This is the specification of location where the item is to be used or to be held for later use.
c. WHEN is it needed? This is the specification of time, either by calendar date or in relation to a proposed time of use.
d. HOW is it needed? This is the specification of the condition of the item in relation to its state at time of use. It covers questions of packaging, assembly, modification, and adjustment.
e. WHY is it needed? This is the specification of the use to which the item is to be put and thereby establishes the relative importance of the item.
f. HOW MUCH is needed? This is the quantitative determination as conditioned by the specifications enumerated above.

The first five of these integrants can be directly derived from the Elements of Strategic Decision but only insofar as they pertain to these elements. That which pertains to the Elements of Logistics Decision can be determined only when the volume of materials to be handled for the Elements of Strategic Decision has been calculated.

In short, requirements determination can be accomplished generally by following a procedure consisting of:

a. Analyzing the stipulated Elements of Strategic Decision to determine the first five integrants.
b. Analyzing these integrants to determine the "how much."
c. Based on a and b design four logistics systems, i.e., the Elements of Logistics Decision.
d. Analyze the Elements of Logistics Decision to determine the first five integrants pertaining thereto.
e. Analyze these to determine the "how much" required for the logistics systems.
f. Combine b and e to determine the total.
g. Adjust the logistics systems to handle the resultant total.

It should be clear from the foregoing that the critical calculation is the "how much," both that which pertains to the strategic systems and that which pertains to the logistics systems. There is no single area in the whole complexity of military planning where the inexactness of the military art is so apparent and so often so embarrassing.

"How much," as the phrase implies, is a quantitative representation that may be expressed either in standard forms of measurement or in forms peculiar to the military as "days of supply" or "units of fire." The term "days of supply" is a useful measurement form because it purports to express quantity in terms of operational effectiveness for a specified time period. However, it has glaring weaknesses in that it can effectively conceal miscalculations of types of materiel within a classification. For example, thirty days supply of ammunition may be authorized and on hand yet contain none of a particular kind of ammunition.

The underlying philosophy of the military as regards the "how much," has been that it is better to have something and not need it than it is to need something and not have it. Every logistician, since the dawn of warfare, has accepted the venial sin of surplus to avoid the mortal sin of shortage. Consequently, the likelihood of "too much" in military stocks will persist as long as there is a likelihood of "not enough."

Since the basic reason for stocking materiel is to preclude exhaustion and since "too much" does at least preclude exhaustion, there is little motivation for seeking a method to determine what is "just enough." Instead the logistician will either rationalize his "too much" and defy the critic to prove him wrong or will, as a last resort, conceal the very existence of the surplus.

Therefore, no method of requirements determination, regardless of personnel or organization, can overcome the deliberate tendency to err in the direction of "too much," until there is some absolution to the mortal sin of shortage.

The concept of risk and hazard is offered as a means of improving methodology in determining the all-important "how much." If such a concept does not provide absolution it at least provides commiseration.

Risk is defined as the probability of an event occurring.

Hazard is defined as the degree of seriousness should the event occur.

In terms of supply requirements, risk is the measure of probability of running out and hazard is the measure of how serious is it if running out does occur.

Risk is a function of the causative agent. Hazard is an expression of military judgment of the relative importance of the item in terms of operational effectiveness. Assuming equal levels of supply of aviation fuel and Kleenex at an overseas base, the risk of exhaustion due to enemy action in interrupting normal flow of materiel to the base would be the same. The difference in hazard is obviously considerable.

The determination of risk requires specification of the various causative agents and their analysis by existing statistical technique. This would include such things as enemy action, "life of type," pilferage, damage in transit, deterioration in storage and damage from use.

The determination of hazard requires the exercise of military judgment on the relationship of the item to operational effectiveness in order to determine how long the using agency can be without the particular item before an unacceptable reduction in operational effectiveness takes place. Naturally, if the item has no bearing on operational effectiveness at all, it has no place in the inventory, since, as stated earlier, the prime purpose of logistics support is operational effectiveness.

Using aviation fuel and Kleenex as examples again, the military cannot afford, even momentarily, any exhaustion of aviation fuel without experiencing an immediate and serious decline of operational effectiveness. Yet, the absence of Kleenex, even for prolonged periods, would be at the worst an inconvenience.

Accordingly, there is no basis in reason for equal standards for the supply of these items. Aviation fuel is, then, a high hazard item and Kleenex, perforce, of low hazard.

An analysis of the hundreds of thousands of items in the military inventory would probably reveal that the great majority of these items are of relatively low hazard.

It would then appear that provision to avoid exhaustion of most of the military inventory decides more than any other single factor the computation of "how much." Yet, acceptance of temporary deprival of low hazard items would reduce the total need for stocks considerably and thus would enable the greater allocation of effort and higher margins of safety to high hazard items within given economic limitations.

A properly designed logistics system enables the systematic satisfaction of requirements and can reduce the likelihood of exhaustion for unacceptably long periods of time by expeditious replenishment without necessitating large stockpiles. A smaller volume in a logistics system of given capacity would naturally decrease the time involved from source to user.

The control of production, procurement, transportation, and distribution is pointed at achieving a never-ending flow of materiel. This is a splendid goal but if measured against a requirement for operational effectiveness in terms of risk and hazard, would prove to be unnecessarily refined and overly costly except for high hazard items.

An analysis of the military inventory in terms of risk and hazard could well demonstrate the futility of central inventory control of low hazard items. Central control of the production, procurement, transportation and distribution of materiel has been justified on grounds of efficiency and economy. Its argument for efficiency has been that such control enables the logistics system to properly respond to strategic need. Its argument for economy has been that volume buying means a lower unit cost.

Actually the best argument for central inventory control is that it militates against exhaustion. Accordingly, it is completely justified for high hazard items since exhaustion of these items is militarily unacceptable. If temporary deprival of low hazard items is acceptable, central inventory control is justifiable only if it is economical.

There is a fallacy in the argument that central procurement is necessarily cheaper. Low hazard items, in large measure, correspond to items having civilian as well as military use. A large purchase of such items tends to firm the market. Secondly, any savings in administrative costs incident to contract negotiation can be quickly dissipated in increased transportation and distribution costs.

Classic examples of this are replete in such centrally procured items as food and solid fuels. In many instances, although payment to producer was at bargain rates, the total cost to the government exceeded local retail costs when transportation, distribution, and administrative costs were added.

Therefore, central control of inventory is justified only when the items are of high hazard and require urgent safeguards against exhaustion or are of a type without ordinary civilian use and require military specifications and standards.

It would logically follow that low hazard items can be stricken from the controlled inventory entirely and their procurement be a function directly of the using agency using funds budgeted for that purpose. This would offer the collateral gain of considerably reducing the size and complexity of the four logistics systems and would enable the using agency to acquire that which best satisfies its own peculiar requirement. The collection of the cash discount for speedy payment customary in normal commercial practice would probably far exceed any collective increase in administrative costs of such purchases.

Further, the elimination of these items from the controlled inventory would mean the abandonment of the vast and complex task of preparing and maintaining specifications and standards. Quality control of those items as exercised by normal civilian competitive practice certainly should suffice for military purposes. The military can choose from a wide variety of fountain pens already on the market. It need not design its own.

It may be argued that procurement is a specialized function requiring skills not available to the using agency and that the decentralization of this function would result in more, not less error in this field. It may be that the risk of error increases, but certainly the hazard decreases. Further, rating the commander of the using agency on his effectiveness in this field will result in rapid skill acquisition.

Certain variations in local procurement procedures would be necessary to cover such military activities as isolated bases and ships at sea. However, the supplying agency can take care of this requirement within the specific request of the commanders of these activities and within the funds budgeted for their use.

It must also be noted that the greatly increased local purchase activity that would result from these proposals is entirely consistent with the widespread development of financial management techniques in the Department of Defense.

The terms risk and hazard have utility other than in supply. They are equally applicable as measurement standards for maintenance, transportation, and construction.

Maintenance planning under peace and war conditions presents entirely different problems particularly when new equipment, untested under combat conditions, is involved. The problem of repair of battle damage to equipment is one of tremendous significance. Risk and hazard calculations would be very useful in this field in determining maintenance standards with its consequent bearing on supply requirements.

Transportation planning can also be conducted effectively under a risk-hazard approach by its bearing on such problems as "peak load," and interruption due to enemy action.

In its application to construction, risk and hazard would be of particular significance under the conditions of modern warfare where such pressing problems as dispersal and protective construction await solution.

Whereas the so-called "calculated risk" in military jargon usually means that the only "calculation" made was that a risk does exist, the techniques of risk determination are in being and require only application.

A technique for hazard determination does not yet exist but should offer no great problem since the relative hazard of most items should be readily apparent from an examination of the first five integrants mentioned before.

Having determined risk and hazard, the obvious next step is their reduction if such reduction is indicated. The relationship between risk and hazard serves as a guide to identifying those areas that require definitive action. Clearly, items of high risk and high hazard require immediate action and correspondingly, items of low risk and low hazard can be ignored.

Reduction of risk involves action affecting the causative agent. Reduction of hazard involves action affecting the item itself. If the risk is enemy attack on a supply point, the causative agent is the enemy force. The reduction of this risk requires active defensive measures and thereby indicates another example of the influence of logistics on strategy. The reduction of the hazard of such an attack involves the security of the supply items and therefore such passive defensive measures as dispersal and protective construction would be indicated.

The foregoing discussion by its very nature is tentative and hypothetical. It is offered as such as a basis for the development of logistics theory. The absence of developed, tested, and accepted logistics theory is, of course, the principal reason why the requirements determiner cannot prove the validity of his calculations. Until sound logistics theory is developed, unsound, erratic, and uneconomical logistics practices will continue to plague the military regardless of any organization changes. In fact, organization for logistics is, in a sense, logistics practice and should itself be based on logistics theory.

The farther logistics practice is removed from military operations, the greater the possibility of erroneously gauging the relationship of logistics to strategy. This risks the neglect of the purpose of logistics which is to contribute to operational effectiveness. The hazard of such neglect is unacceptably high.

* * *

COMMENTS ON THE MORGENSTERN MODEL

Commander J. C. Busby
Supply Corps, U. S. Navy

The author presents some ideas, stimulated by a recent article in this Quarterly, on management of an operating supply system.

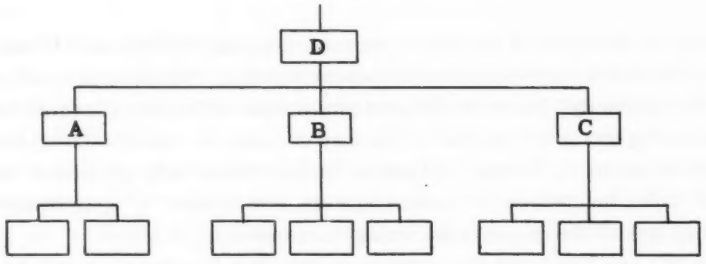
In a recent article[1] O. Morgenstern laid the theoretical foundation for what may become an important tool in improving our basic understanding of military supply systems. This increased understanding has already led to significant changes in operations within the Navy Supply System, and no doubt further theoretical exploration will prove fruitful. I propose in this paper to comment on the importance of Dr. Morgenstern's work in orienting military value judgments. Also I would like to elaborate on the basic model of an overseas military supply system and point out how the more elaborate model can serve as a powerful tool in the management of an operating supply system.

MODELS AND VALUE JUDGMENTS

It is commonplace knowledge that human decisions in the management of a military supply system are largely based on conceptual models. Reference to the organization of the Navy Supply Center Oakland immediately evokes the brain picture of a chart with many little boxes connected by lines. Reference to the continental U.S. supply distribution system evokes a picture of liquid flowing through pipelines which connect various tanks.

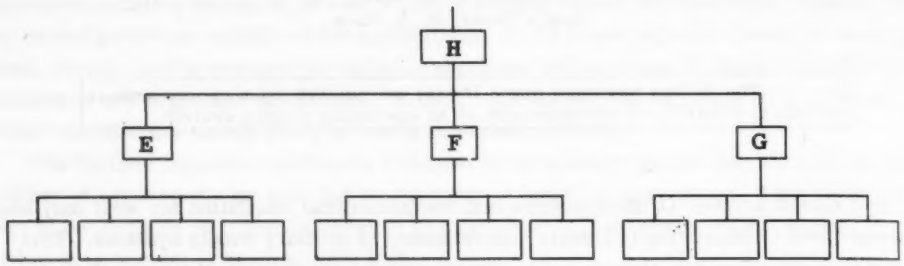
The difficulty with many of the widely used conceptual models is that they tend to force the man using them to employ value judgments which are poorly related to the problem at hand.

I remember very well an inspection conducted at a large supply activity several years ago. In examining the organization of the activity, the inspecting party noticed that at one place on the charts there appeared blocks like this:



[1] Morgenstern, O., "Consistency Problems in the Military Supply System." Naval Research Logistics Quarterly, Vol. I (1954), pp 265-281.

whereas throughout the rest of the charts blocks looked like this:



The inspecting party concluded that the supervisor occupying block "A" was unnecessary since only two persons were supervised, and recommended a reorganization which was eventually put into effect at a distinct detriment to the Navy.

Why did this occur? The model of the organization—the chart with boxes and lines—has built into it a false value judgement: symmetry. We conceive the organization as a pyramid, each block being supported by several below it. When the model shows up some deviation from this symmetrical structure, we take a close look at the segment which is "out of step." And even the best intentioned, most open-minded inspection officer imaginable inevitably listens with only half an ear to the protests that "special circumstances" warrant the nonsymmetrical organization.

The point is this: what does symmetry of boxes on a chart have to do with efficiency or effectiveness of an organization?

There are many similar examples of conceptual models giving rise to value judgements which are either vaguely or not at all related to the problem being studied.

A particularly dangerous tendency now prevalent is to consider military problems from the viewpoints of "sound business administration," "cost consciousness," "efficiency" and so on.[2] These conceptual models have the built in value judgement of "efficiency of employment of resources." Somewhere along the way the value judgement of "military effectiveness" gets lost.

A group of management engineers recommends that Air Station "A" can save 20% of its storage space by switching to a system of locator storage. This fits in readily with the value judgements of "efficiency" based on the conceptual model of a pair of scales with one pan full of dollars balancing the other pan full of quantity of material issued. Then along comes a one hour notice of an enemy air attack, and before inexperienced duty personnel can locate needed oxygen masks under the "efficient" locator system, Air Station "A" gets destroyed because our planes could not get off the ground soon enough to protect it.

Those who are faced with the awful responsibility of achieving the delicate balance between morale of men and cost of material, between present-day system efficiency and wartime system reliability, between public acceptance and military need, et cetera ad infinitum, have achieved through years of experience a certain intuitive manner of looking at problems with value judgements based on military effectiveness. The Nation will be well served if the conceptual models used by these men can be set down, clarified, improved, and publicized

[2] For an analysis by a seasoned observer see Hanson Baldwin's article "Pentagon and Economy" in the 4 June 1955 NEW YORK TIMES.

sufficiently to influence that large group of personnel, military and civilian, who have fallen into the habit of paying lip service to "military capability" and who then proceed to evaluate all problems from the standpoint of a grocery store owner trying to cut costs.

THE MORGENSTERN MODEL

The great benefit of Dr. Morgenstern's model of a military supply system is that it is readily adaptable to manipulation using value judgements based on military effectiveness. In essence, an overseas military supply system is described as consisting of a supply center in the United States supporting one or more overseas bases (or supply ships). These bases, in turn, each support one or more activities.

In the course of events, an overseas base may be unable to supply one of its supported activities with a specific item at the time required. Such a situation is described as a "state of deficiency," and it is the duty of the overseas base to take measures to avoid a state of deficiency, or to minimize the implied losses when it cannot be avoided.

(Note the built in value judgement of "military effectiveness" under this concept. A less experienced investigator might easily have been tempted to imply that the duty of the overseas base was to supply its customers at a minimum investment in men and material.)

There are at least two kinds of states of deficiency:

(1) ordinary, occurring when the base is unable to fill a requirement for an item which it normally carries on hand. This is commonly known as an "NIS" or "not in stock" deficiency.

(2) special, occurring when the base is unable to fill a requirement for an item not normally carried. This is commonly known as an "NC" or "not carried" deficiency.

States of deficiency can be remedied by two means:

(a) Locally (i.e., at the overseas base):
  (1) by repair
  (2) by substitution
  (3) by cannibalization
  (4) by provisioning:
    (a) purchase
    (b) confiscation

(b) Nonlocal:
  (1) by requisition on the supply center.

In order for the supply center to have some guidelines in planning its internal operations, Dr. Morgenstern proposes that each incoming requisition from an overseas base should contain two pieces of information:

(a) expected date of the state of deficiency in case the current order should remain unfulfilled as of that date;

(b) priority, or "ordering of importance of the expected deficiency to the particular overseas base."

The foregoing summarizes the main features of the Morgenstern model which concern us at this time.

A major purpose of the original model was to demonstrate that the Navy Supply System, as it was then constituted, was not internally consistent. An overseas base was required to put on its requisitions not the expected date of the state of deficiency, but rather a "deadline delivery date" based on its estimate of processing time at the supply center, shipping time, and other factors. Since the overseas base could not possibly know this information on a

current and accurate basis, the result was that operations at the supply center and of the transportation agencies were governed by guidelines which did not reflect optimum military effectiveness. A motion picture projector with a "guesstimated" DDD two days hence could easily squeeze off the airplane a part for an inoperative 5" gun which carried an equally "guesstimated" DDD of three days hence.

The original model served its purpose of demonstrating inconsistency very well indeed. But there is no reason why the model should not be elaborated and made to serve an even wider purpose. It is this task which shall next be addressed. For convenience, the major points of elaboration will be given paragraph numbers.

ELABORATION OF THE MODEL AND SOME PRACTICAL APPLICATIONS

1. Each activity supported by an overseas base maintains stocks for its own needs. As with the base, a state of deficiency will occasionally arise which cannot be remedied locally at the activity. Statistical data on such states of deficiency can be collected by analysis of the requisitions received by the overseas base and can provide a useful management tool as explained in paragraph 2.

Recognition of the fact that duplication of stocks exists at the several activities, the overseas base, and the supply center poses the problem: "how should levels be set at each of the three supply echelons?" The higher in the echelon that stocks are held, the less total stocks are required due to pooling of safety levels. The lower in the echelon they are held, the less total effort will be required for processing papers and shipping material since there will be less frequent re-ordering by activities. More important, states of deficiency at the activity level (where they are most harmful to military effectiveness) will occur less often when stocks are held at the local activity in lieu of at the overseas base or the supply center. Furthermore, the system as a whole is less vulnerable to enemy attack or other disruption since stocks are dispersed and since activities are more self sufficient. In short, greatest military effectiveness and least cost in processing papers and shipping materials is achieved by keeping activity stock levels high; economy in total inventory investment is achieved by keeping activity stock levels low. The supply system manager can obtain valuable control data under this conceptual model by analyzing the stock levels at activities, overseas bases, and the supply center. There is no doubt that a point of optimum economy in placement of stocks can be found, and that deviation from this point can be authorized depending on the price we are willing to pay for improved military capability. There are some observers who believe that the Navy has its present stocks poorly located from the standpoint of both dollar economy and military effectiveness.[3]

2. The two classes of states of deficiency established in the original model should be expanded and put to practical use. A convenient classification is:

(a) "NIS" or "not in stock," occurring when a demand cannot be filled for an item normally carried. The two types are:

(1) "NIS," occurring when the stock card shows none on hand, hence no attempt is made to find the item.

(2) "NIS-WR," a "warehouse reject" occurring when the stock card shows a balance on hand, but the item cannot be located in the warehouse.

(b) "NC," or "not carried," occurring when a demand cannot be filled because the item is not normally carried on hand. The three types are:

---

[3]See, for example, Lynch, Captain Charles F., "Notes on Applied Analytical Logistics in the Navy." NAVAL RESEARCH LOGISTICS QUARTERLY, Vol. 1 (1954), pp. 90-102.

(1) "NC," occurring when the item is "not carried" and no action has been taken to carry it, even though it is stocked elsewhere in the supply system.

(2) "NCL," or "not carried but has been added to the load list," occurring when the item has not been carried in the past, but action has been taken to carry it (i.e., it has been "added to the load list") although the initial increment of stocks has not arrived.

(3) "NSI," or "not stocked in the system," occurring when a demand is received for an item which is not carried anywhere in the supply system. (This occasionally happens when replacement parts are needed for nonstandard equipment.)

(c) "NID" or "not identified," occurring when a demand is received for an item which cannot be identified with a standard Navy stock number and which cannot be positively identified as an item not stocked in the system (NSI).

Practical uses of the above classification are self-evident. The supply manager can gauge the internal efficiency of activities by the NIS and NID rates, and particularly by the NIS-WR rate. The NID rate also reflects the adequacy of identification media provided by higher authority, and the state of training and supply discipline of end users.

In a supply system where the stock levels at an activity are set by higher authority, the NIS rate also reflects the depth adequacy of the levels. The NC rates reflect the accuracy with which range requirements have been forecast, and the NCL rates the effectiveness of the mechanics for putting into effect such forecasts. The NSI rate indicates the adequacy of the stocks in the entire supply system to meet the range of requirements of end users.

Although this classification of states of deficiency provides a tool for gauging the military effectiveness of various components in the supply system, it cannot be used blindly. Considerable experience is required to evaluate the "goodness" or "badness" of any given rate or even a trend.

3. The original model should be expanded to show that states of deficiency can be remedied locally by any of the following means:

(a) by repair or manufacture or cannibalization (R)

(b) by substitution (S)

(c) by provisioning (P)

(d) by redistribution of stocks—e.g., by forwarding the demand to some other activity in the area known to have the item on hand. (F)

As a practical matter, the supply officer of an activity or overseas base will seldom have direct charge of repair, manufacture, or cannibalization and it is of no particular concern which of these three methods is used to remedy a deficiency. Cannibalization usually consists in removing parts of an already inoperative, or perhaps scrapped, piece of equipment which is not carried as a stock item. Furthermore, the dividing line between repair of an item and manufacture of a new one is frequently blurred. For these reasons, it is sufficient for the model to indicate that a potential source for remedying a state of deficiency is a local repair activity which either manufactures a new item, repairs the old, or perhaps cannibalizes some existing piece of equipment.

In the Navy Supply System, "R" remedies (that is, filling demands by local repair, manufacture or cannibalization) are of great importance. There are few items which cannot be manufactured or repaired afloat, and the supply system manager should be kept continually informed of the degree to which states of deficiency are remedied by this technique. A lack of statistical data can lead to decisions which will reduce military effectiveness by eliminating local incentives to maintain a repair capability. It has been my personal observation that the

average Chief Engineer of a ship is not nearly so prone nowadays to manufacture a repair part as he was during World War II. He is reasonably sure he will get his part within a few weeks nowadays, whereas during the war requisitions sometimes never got filled. A system of gathering data on this problem at the type commander or fleet level would have shown the trend years ago and perhaps have resulted in corrective action.

Turning to substitution as a means of remedying states of deficiency, it appears that two elements are involved: recognition and exercise.

To recognize the possibility of substitution requires only a knowledge of technological interchangeability of the items and their approximate relative stock positions. Technological interchangeability may be of commonplace knowledge (e.g., canned ham may be used in place of smoked ham); or it may require detailed knowledge of the intended end-use (e.g., cabinet fittings and hardware of one type may or may not be usable in place of those of another type); or it may require intimate technical knowledge of the items themselves (e.g., when one electronic tube can do for another).

Note that knowledge of relative stock positions need not be exact. The mere fact that an overseas base "ought" to have the substitute item in stock may be adequate to enable a consuming activity to recognize a substitution possibility and to pass such information on to the overseas base.

To exercise the possibility of substitution requires in addition, and as a minimum, only a more detailed knowledge of relative stock positions. For optimum operation, however, exercise of substitution also requires knowledge of the prospective quantity and time of inflow (or "resupply") and "outflow" (or "demand") of the items involved; the military value or importance of such prospective demand; and the intrinsic value of the items involved (which, in the absence of more appropriate military or economic value measurements, must usually be measured in terms of catalog price or dollars).

The quantity of substitution that takes place in supply system can provide important measures of operating efficiency and of the states of information available at various levels. The types of substitution that take place in a supply system can provide equally important measures of range adequacy (i.e., suitability of the range of inventory items normally carried in stock to meet the demands placed on that stock), and of economic efficiency (as measured by optimum use of available resources).

During war and other times of critical shortages, intensive efforts are made by both consuming activities and higher supply echelons to increase states of substitutability information available at various levels in the system. Usually this will result in an increase in the quantity of substitution that takes place, but not necessarily in the economic efficiency of the supply system.

Local remedy by provisioning may take the form of purchase or, in wartime, of confiscation. While these two methods are of theoretical importance, it is not probable that the supply manager need concern himself with any breakdown between the two in either peacetime or such wartime practices of the U.S. military forces as are presently envisioned. Hence the model may be simplified by combining both confiscation and purchase (which includes such present practices as obtaining milk from recombining plants and vegetables from hydroponic farms) into the single category "P" or local provisioning.

Local remedy by redistribution of stocks is possible because in practice stocks at supported activities are ordered in "economic" quantities, which frequently means quantities in excess of immediate needs. For example, a given type of set screw might well be of great

military importance but have an extremely low issue rate. Supported activities will naturally order quantities in excess of their actual requirements because the item is inexpensive and occupies very little space, yet is too important to risk not having on hand a few extra ones to meet unforeseen contingencies. The subject of economic ordering quantities has been pursued at length by Whitin[4] and others, and it is sufficient to say here that the possibility of redistribution of stocks does not necessarily imply a defect in the supply system; rather it recognizes that material placed at given locations under one set of military value assumptions need not necessarily be at the optimum places should the value assumptions be changed.

To summarize: our conceptual model shows that it is the duty of the supply echelons to avoid states of deficiency, or to minimize the implied losses when avoidance is impossible. There are four local remedies for states of deficiency: repair, substitution, provisioning, and redistribution.

From the standpoint of military effectiveness, the supply system manager should constantly review the extent to which the four possible remedies are in fact used, and should be prepared to take administrative action when it appears that loss of military readiness is occurring through failure to maintain an exploitation capability in any one of the four channels.

4. Finally, I propose to modify the original Morgenstern model by establishing a more realistic system to guide priority of effort of the supply center and transportation agencies.

One of the basic problems in any supply system is that, for sound military reasons, all the organizations which must work together to supply material to overseas bases are not under a commander low enough in the echelon to exercise day-to-day detailed control of operations. This situation is the same as exists in civilian life where General Motors, for example, is dependent on its subcontractors, the communications companies, the U.S. Mail Service, transportation agencies, and various others to maintain an effective resupply system to its numerous plants.

Therefore, there arise everyday problems whose implications are clear from such typical statements as: "If the overseas bases would only send in their requisitions by dispatch instead of mail, we could get the material to them on time."

"If the supply center didn't take so long to process a requisition, we wouldn't have to use all this air transportation."

"If the transportation people would only use a little judgment in diverting ships and planes and routing freight, these delays in receipt of material would never occur."

In order to analyze the method by which an overseas base obtains material, let us establish a conceptual model of an assembly line system consisting of three sections as shown in Figure 1.

The first section governs the preparation of a requisition at the overseas base and its transmittal to the supply center. In practice, this section can run at two speeds: (1) fast (expedited preparation of the requisition and message transmission); or (2) routine (normal preparation of the requisition and mail transmission). Selection of the speed is entirely under control of the overseas base.

The second section governs the processing of the requisition and issue of material at the supply center. In practice, this section can run at three speeds: (1) super-speed (the "walk-through" or "issue on requisition" technique by which all possible short cuts are taken

---

[4]Whitin, T. M., THEORY OF INVENTORY MANAGEMENT. Princeton: Princeton University Press, 1953.

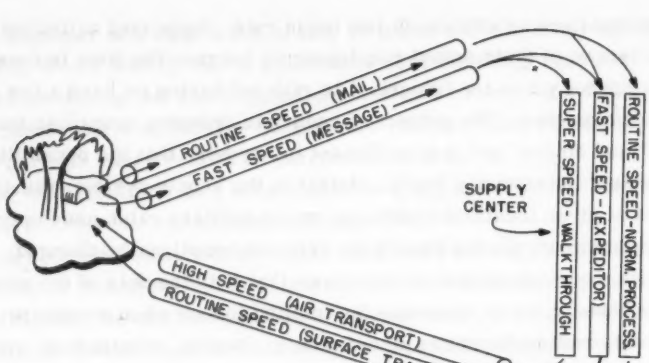


Figure 1 - Assembly line model of overseas base replenishment

and the material issued as rapidly as possible); (2) fast (the so-called "expeditor processing" by either a separate group of clerks or by the same group which handles routine requisitions but which operate under instruction to handle "expeditor requisitions" ahead of routine ones); or (3) routine (the normal processing technique which is designed to achieve a high level of internal efficiency at the supply center).

As a practical matter, it is not feasible to operate any one of the three channels in the supply center section of the line at a speed other than the one for which it is designed. In other words, there is no way a requisition once placed in the "routine" channel can be speeded up except at an inordinate cost and most probably with little resulting speed-up. The reason for this can be readily seen when one reflects on the tremendous number of operations involved in issuing a single screwdriver. These operations are performed on a mass production basis by many persons in various separate organizational sections, and any attempt to locate a single piece of paper once it is in the assembly line system is a time-consuming task. An attempt to locate the several pieces of paper involved in processing a single requisition is usually so time-consuming that it would be quicker to let events take their normal course and wait until the completed product emerges from the assembly line.

The final section of the assembly line, governing transportation of the material, operates at two main speeds (surface or air transportation) either of which may usually be selected in advance by either the overseas base or the supply center. Within the two main speed brackets, the assembly line sometimes has a further range of speeds which are controlled by the transportation agency. Thus, a package destined for air flight can be put on a plane today or wait until tomorrow depending on the priority it bears relative to competing passenger and freight traffic.

Having established the basic model of an assembly line system by which bases are replenished, we next turn to the problem of how to optimize the operation of the system.

First, assume that the most economical operation takes place when each section of the assembly line operates at its slowest speed. That is, mail transmission is assumed cheaper than message; routine issue of material cheaper than expeditor issue; surface transportation cheaper than air. Second, assume it is desirable from the standpoint of military effectiveness

to process as many requisitions as possible by the most economical means—that is, at the slowest speed on each section of the assembly line. In other words, from the standpoint of using our limited resources for greatest military effectiveness, we prefer to keep supply matters off the radio circuits; use minimum personnel in supply centers; and to use surface in preference to air transportation.

Under these assumptions, we purposely stock the overseas bases in a manner which will permit them to replenish their stocks under ordinary circumstances by using only the slowest assembly line speeds. This requires either a continuation of the present Pacific system of "scheduled replenishment," or else some other system which will guarantee a known and rigidly enforced lead time.

In a typical situation, overseas base "Y" will review all stock cards in class Q25 between the 10th and 14th of each month, submit requisitions by the 15th, which will be received and put in process by the supply center on the 25th, scheduled to be loaded aboard a "pipeline ship" by the 15th of the following month, and received and taken up in stock at base "Y" by the 30th. This fairly rigid schedule enables accurate prediction of lead time (which is 45 days from date of submission of the requisition in the example given).

Hence, for the great majority of requisitions, adherence to the established schedule of operations will be sufficient to guide operations at the supply center and of the transportation agencies in a manner which will effectively prevent the occurrence of states of deficiency.

There will, however, be a minority of requisitions which must be processed more rapidly than the slowest speed assembly line operates. They can originate from only two basic causes:

(1) "range inadequacy"—which occurs when a demand occurs for an item not carried by overseas base; or

(2) "depth inadequacy"—which occurs when an emergency requisition must be submitted to keep from running out of stock normally carried at the overseas base.

As a practical matter, the supply system manager knows fairly accurately what NIS rate any given overseas activity should be able to achieve under his allotted stock levels. Hence, an excessive number of requisitions submitted for emergency replenishment of normally carried stocks will indicate either poor local stock control procedures or else an abnormal demand pattern due to causes not known or not yet analyzed by the supply system manager.

Demands due to range inadequacy should occur in such amount as the overall supply system is designed for. That is, stocks at overseas bases are ordinarily established on a basis which should satisfy some given percentage of the demands which will be placed on them (say 90 or 95%), and it is to be expected that the remaining demands will be passed back to the next higher echelon of supply (in this case to the supply center). An excessive number of requisitions due to range inadequacy usually indicates lack of effort to exploit three of the four local remedies for states of deficiency (repair, substitution, and provisioning).

From the foregoing it is seen that the conceptual model establishes a supply system designed to process the majority of requisitions by the slowest speed assembly line, and a minority of requisitions by higher speed lines. Furthermore, the amount of such minority requisitions can be estimated in advance by the supply manager, and arrangements made to provide communications channels, air transportation, and additional people at the supply center

to handle this minority amount.[5] Should minority requisitions exceed the estimated amount, some malfunctioning of the system exists and the source can be readily determined. Administrative action would then be taken to correct the situation.

Before going further, it should be pointed out that the foregoing model has been simplified to illustrate certain basic principles. For example, considerable effort is expended to have only the most "critical" or militarily useful items in the forward area, reserving the luxury items for handling thru the next higher supply echelon. Stock levels are not uniform by any means, but dependent on such diverse factors as cost, volume, perishability, military essentiality, and so on. Finally, lest there be any who feel that an exact mathematical treatment of the problem is feasible, I should point out that in reality there are so many intangible factors such as inability to estimate enemy action, lack of precision in forecasting our own tempo of activity, arbitrary budgetary decisions by higher authority, et cetera ad nauseum, that a great deal of professional experience and judgement is required to interpret and act in any given situation.

Returning to the assembly line model, the problem of guiding the operations of the supply system to ensure that anticipated losses due to unavoidable states of deficiency are minimized boils down to putting the minority known amount of requisitions on one of the higher speed assembly lines in accordance with the relative military value of the requisitions.

A convenient method of doing this is to include in each requisition the date the material is required or "DMR." This date is the same as the "commencement of a state of deficiency" as given in the original Morgenstern paper. However, the term "DMR" is a simpler expression than "state of deficiency" and will probably be more readily and correctly understood by the thousands of people who must eventually be trained to use it.

In addition, the "DMR" should be followed by a suffix "C" or "D" to indicate that, if the material is not received by that date, the need for it will continue or decrease. To illustrate:

"DMRC 12 Sept" means "material is required by 12 September and the need for it continues thereafter." A typical example would be a replacement typewriter or motion picture projector for an overseas base, or boiler tubes to repair a ship.

"DMRD 12 Sept" means "material is required by 12 September, but if that date cannot be met the need will become less urgent, although the material is still desired." Typical examples would be allowance list replenishment items for a ship due to sail on 12 September, or a new drug particularly desired to treat a patient prior to his scheduled evacuation on 12 September.

It is obvious that such a coding will usually enable proper selection of speed at each section of the assembly line under normal operations when the capacity of the higher speed channels is in fair balance with the amount of requisitions received which bear a DMR.[6]

Under wartime conditions, the capacity of the higher speed channels can easily get out of balance with the volume of requisitions which must be processed on them. Diversion of aircraft for special lift (as occurred during the Berlin Blockade for example) can reduce the

---

[5] A description of the techniques by which the supply manager determines range and depth of stocks at overseas bases, and balances these against military and economic costs of running high speed assembly line sections is too complex to be covered here.

[6] Note that it is possible to have a DMR later than the date on which material would normally be received via the slowest speed channel. For example, a ship repair facility might well order material for overhaul of a ship six months in advance of the time when it will be required. Operating rules can be established to handle such cases in any of several ways, and there is no advantage in pursuing the matter further here.

capacity of the high speed channel in the third section of the assembly line to near zero. Hence, some method is required to distinguish which of the minority requisitions are least important so that they may be pushed back to a slower speed assembly line channel. Obviously, this means some method must be devised to classify minority requisitions by their over-all military importance (not their importance to a particular overseas base). To accomplish this task, we introduce a "Priority Reference Standard," or "PRS," which serves as a basis for identifying the military importance of various end uses to which requisitioned items may be put. For example, we might use a two-value PRS such as:

"EMERGENCY"---assignable to any item necessary to carry out an assigned mission, or to any item whose lack would impose serious personnel hazard or discomfort.

"ROUTINE"-------assignable to all other items.

Or we might use a three valued PRS:

"A"----------assignable to any item necessary to carry out an assigned combat mission.

"B"----------assignable to any item necessary to carry out an assigned mission other than combat, or to any item whose lack would impose serious personnel hazard or discomfort, or to any item required for special programs designated by a fleet commander or higher authority.

"C"----------assignable to all other items.

With a PRS as indicated, assignment of priorities is not a matter for local decision, but is rather a question of fact. Interpretation is required, of course; but so is interpretation required of any law or regulation ever issued. For that reason, it is desirable for the PRS to be fairly explicit and to include detailed examples to guide responsible personnel in correctly classifying requisitions into their proper priority category.

To conclude the conceptual model: we have a minority of requisitions which must be processed by high speed channels in the various sections of the assembly line. Each of these requisitions bears a notation showing:

(1) the date the material is required;

(2) whether the need continues or becomes less urgent after that date; and

(3) the relative military importance of the material in accordance with a universally used priority reference standard or "PRS."

Under normal operations, the capacity of the high speed channels is adequate to take care of the minority group of requisitions, hence selection of speed at each section of the assembly line is governed entirely by the date the material is required.

When the volume of requisitions being processed on the high speed channels exceeds the available capacity, the supply system manager can arbitrarily order a diversion of the least important requisitions into slower speed channels. These least important requisitions are easily recognized because all requisitions originally destined for other than routine processing must carry a priority assigned on the basis of the ultimate military value of the item as determined from a universally known and used priority reference standard.

FINAL REMARKS

The foregoing proposed modifications to the Morgenstern model are not intended to establish a "true" picture of military supply system operations. No model can ever be complete. By using the concepts advanced, however, very practical working tools can be developed

from the available statistical data. Furthermore, the concepts can serve as a basis for obtaining a common understanding and frame of reference for making day to day decisions in supply management.

I would like to emphasize again that the model can assist decisions by properly orienting military value judgements—but it can not replace professional judgement and experience which are the most important ingredients in the decision process.

* * *

ON THE MEANING AND USE OF A CAPACITY CONCEPT[1]

Edward Zabel
Princeton University

The author surveys and interprets previous attempts to define and measure capacity, and analyzes the use of the capacity concept in economic models.

Part I
MAXIMUM OUTPUT

INTRODUCTION

Despite the fact that in recent years studies on capacity have appeared more and more frequently,[2] it is still common to treat this concept gingerly. A favorite device of some authors is to clothe the term in quotation marks, revealing, presumably, an awareness of its unmistakable ambiguity. This ambiguity is perhaps the essential feature of the problem; the term, capacity, can mean almost anything, and it can mean virtually nothing. Yet the notion of a capacity can be useful if formulated clearly. The first requisite is that objectives and conditions be explicitly specified for each study. For example, to select some number or array of numbers to represent the capacity of the steel industry or a naval supply depot is meaningless without specifying whether we consider the stock of capital goods to be fixed or variable, permit labor to work overtime, measure capacity for an instantaneous or finite period of time, and so on. A point which may make the study of capacity difficult is that we are almost always interested in measurements in a system of interdependent organizations; i.e., a system in which the activities of any organization are related to the activities of other organizations. In measuring capacity it is well to remember that resources have alternative uses and are better allocated to some employments than others.

Conceptually, it seem plausible to classify the present approaches to the study of capacity into two related categories. The first approach concerns the maximum output producible during some period of time under given conditions of demand and costs. With this approach it is conceivable to have a sizeable number of capacity measures, depending only on the number of possible combinations of conditions. The second approach involves a balance among a number of variables. For example, for an economic organization, capacity may be defined as the optimum level of stocks of goods which are used to produce the output. This is determined by balancing the various costs involved. When the proper balance does not prevail, excess or under-capacity exists and adjustments in the present stocks are profitable.

---

[1]The research for this paper was carried out under Office of Naval Research Contract N6onr-27009. The author is indebted to Professor Oskar Morgenstern, Professor Jesse Markham, Dr. Martin Shubik and Mr. William L. Baldwin of Princeton University for helpful suggestions and criticisms.

[2]Consult Annotated Bibliography on Industry Capacity, U.S. Department of Interior, Bureau of Mines, Inter-Industry Analysis Branch, Item 5, May 9, 1952; and J. N. Behrman, "Notes on the Concept of Capacity and the Measurement of Capital Intensity," Logistics Papers, George Washington University, 1952.

In this paper we propose to discuss some conceptual aspects of the capacity problem.[3] For the remainder of Part I we discuss the first approach in more detail and survey an actual attempt to measure the capacity of the United States economy. In Part II we consider the second approach and analyze the use of a capacity concept in investment models in economics.

THE MAXIMUM OUTPUT APPROACH

As we have stated, this approach involves measuring the maximum output which can be produced, given certain conditions and objectives. The crucial point is to specify the conditions and objectives, and only when this is done can we be certain that the concept is unambiguous. It is easy to show why this is necessary by giving a simple example of the measurement of capacity output. With the same example we can also indicate some common measures of capacity. Consider a profit-maximizing bus company which has one bus and two routes; one, say, in the heavily populated section of town and the other in the sparsely populated section. The public service commission has set a fare which is the same for both routes and has decreed that each route must be traveled a minimum number of times each day. At the given fare the demand for transportation services is known for each route. What is the capacity, or maximum output, of the bus company in terms of passengers? It is obvious that there is more than one answer to this question. For example, as a solution we may determine the maximum number of passengers the company can transport per month over both routes after adjustments for normal starting delays, traffic conditions, and maintenance time. The measure is, then, independent of the demand for transportation services and the costs of operation.[4] Apart from being a rough indicator of the size of the company, it is difficult to envisage specific uses for this type of measure. As an alternative answer we may introduce the influence of demand and costs on the number of passengers and then, assuming that the company maximizes profits, we determine the most profitable output, i.e., given the fare and the demand for transportation services, we schedule the bus for the two routes such that no additional gain accrues from some other schedule.[5] We may consider yet another answer. We may measure what happens to the number of passengers when the public service commission changes the fare or the local

---

[3] For the sake of the argument, we assume that the problems of measurement and appropriate data are solved. For an analysis of these difficulties, cf. Edwin G. Nourse and Associates, America's Capacity to Produce, The Brookings Institution, Washington, D. C., 1934, passim; Arthur F. Burns, "The Brookings Inquiry into Income Distribution and Progress," Quarterly Journal of Economics, Vol. L, No. 3, May 1936, pp. 484-489; and J. N. Behrman, op. cit. Additional references are in Bureau of Mines, Inter-Industry Analysis Branch, op. cit.

[4] Industry measures are frequently made under similar conditions and are commonly designated as technological measures of capacity. For definitions of capacity for various industries, cf. Bert G. Hickman, "Cyclical Variations in Capacity and Capacity Utilization in Selected Industries," Conference on Research in Income and Wealth, National Bureau of Economic Research, October 9-11, 1953. A typical example is the following: "Paper 1917-1950. Source: American Paper and Pulp Association, Statistics of the Paper Industry, 1951, p. 53. Unit: Ton. Capacity measured on a tons-per-day basis is multiplied by 310 days to estimate annual capacity. Daily output is calculated for a 24-hour day with an allowance for normal cleanup. Principal Characteristics: Continuous operation. Capacity is defined as the maximum amount of paper that can be produced by a given set of equipment based on normal expectations of grade and weight expected to be produced." p. 5.

[5] Arthur F. Burns seems to have had this type of measure in mind when he listed five propositions for a satisfactory concept of capacity. We quote the second and major proposition. "The second proposition is that 'capacity to produce' is an economic and not a technological problem, the thing to be maximized being human satisfactions. Since productive resources can be assigned to ends of varying urgency, the real problem is what allocation of resources will maximize production in the sense of a sum of satisfactions, not in some technological sense—say, in terms of weight or bulk." Arthur F. Burns, loc. cit., p. 479.

taxi company varies its rates. In other words, we measure the response of output to changes in conditions facing the company.[6]

The first measure is conventional and little more need be said about it, while the latter two measures are solutions to problems of immediate interest. We can generalize these methods: 1) Given the mechanical facilities for an organization, what is the maximum physical output which can be produced with these facilities after allowances for normal interruptions; 2) given a preference function for an organization (the profit function in the example), resources available, prices, and techniques, what is the most profitable output, i.e., the output at the maximum position of the preference function; and 3) assuming that presently the preference function is at its maximum, what is the new most profitable output if conditions change?

While other measures are conceivable, the answers we have proposed to the question of maximum output provide the most common measures of capacity. Aside from industry concepts, studies on capacity are frequently posed in such a way that one of the latter two measures mentioned are demanded as answers. For example, to consider a well-known study, suppose we wish to measure the effective capacity of the economy to produce a single product, or, in other words, the actual output of the product which the economy is capable of producing in some period of time.[7] How is this to be done? We could treat this as a problem of efficient allocation of resources with a given preference function, as in 2), and we would measure the output of the product when there is an optimum distribution of resources in competitive uses. Alternatively, we could treat this as a problem of measuring what happens to output when conditions change, as in 3). Terborgh chose the latter method. We shall briefly summarize his argument.

Terborgh maintains that it is a popular notion that for each commodity produced in the economy there is a definite measurable productive capacity, and this notion, he claims, is derived from an exclusive preoccupation with present mechanical factors of production (such as in 1). Such a concept is very restricted as the possibilities for stepping up output during some period are more numerous than simply increasing the intensity of use of present mechanical facilities. The following are some of these possibilities: obsolete and idle machinery and plant can be reactivated, mechanical facilities elsewhere in the system can be converted to the production of the commodity, the practice of subcontracting to other firms for parts can be used, and new facilities can be constructed and utilized for new production during the period. Since for almost any product adjustments such as the above are possible, it is only in a limited sense that present mechanical facilities set limits to the maximum output.

---

[6]Charles Goodeve distinguishes three types of capacity—a holding capacity, such as the number of people which can be put into a bus; a rate capacity which is similar to the first measure of capacity for the bus company; and an elastic capacity which presumably is similar to the one just mentioned above. He says, "The third type, elastic capacity, has not yet really come into operations, largely because the quantities concerned are very difficult to measure. Prices and incentives are examples of intensive properties or can be defined as such. The supply of a given commodity is said to be elastic and depends upon the price. When this relation is a linear one and both quantities can be measured, then one can define and use an elastic capacity for this system. One can predict that the capacity, in the case of the change of supply in steel scrap in relation to a change in the price, will be very low; in other words, the market is inelastic. On the other hand, the supply of a mineral can commonly go up rapidly with prices, and this system could be considered elastic with a high capacity." Charles Goodeve, "Operations Research as a Science," in Joseph F. McClosky and Florence N. Trefethen, eds., Operations Research for Management, The Johns Hopkins Press, 1954, p. 63.

[7]George Terborgh, "The Problem of Manufacturing Capacity," Federal Reserve Bulletin, July 1940, pp. 639-642.

In addition, there is no a priori reason why mechanical facilities, whether present plants assigned to the product, potential new mechanical facilities, or convertible facilities, should constitute the effectively limiting variables. In increasing output, bottlenecks may as well be caused by shortages of complementary factors, e.g., labor and raw materials. A capacity measure must admit the possibility that any of the factors necessary for the production of a commodity may become limitational at any time. An exhaustive analysis must underline a useful capacity measure showing what will be the limiting variables at any specific time, how adjustments are to be made, and how long it will take. Viewing all these things, rather than have a capacity measure for some period of time, it will probably be necessary to consider the movements of capacity through time. That is to say, we would want a function showing how the instantaneous rate of output should change over time as various adjustments are made.

In brief, if we measure the capacity output for a single commodity, we must allow for all adjustments which may increase output and, since these adjustments may be made at different rates, we should indicate how the capacity output grows over time.

As yet, we have not explicitly mentioned incentives. In practice, a commodity is produced under certain conditions of demand and costs, and its rate of output will vary only if the conditions change. A measure of effective capacity presupposes that we designate these particular conditions. Terborgh maintains that the only valid test of effective capacity is performance under specific circumstances and that there may be as many estimates of capacity as there are combinations of conditions both of supply and demand. The capacity output of a single commodity, then, is dependent, directly or indirectly, on all the conditions of supply and demand prevailing in the economy. For example, it is important to know what is the general level and pattern of production in other parts of the economic system, as this is certain to affect the reaction, say, to an increase in demand for the commodity.

In order to indicate more forcefully the importance of demand and supply conditions or, generally, a value system, suppose we were to attempt to measure the effective capacity for all the commodities produced in the economy. We would be at wit's end unless we were to use a system of prices or allocators of some sort to determine the distribution of resources to alternative uses. Several works have been concerned with the capacity to produce of the United States economy.[8] To illustrate again some of the conceptual pitfalls involved in the measurement of capacity we shall survey the Brookings study in the next section. The study was formulated such that a method similar to 2) or 3) above was needed but the authors failed to recognize this and, consequently, the result, in general, was a failure. However, the problem of measuring the capacity of an economy has never been satisfactorily solved, though the present-day input-output studies may be considered as a step in this direction.

THE BROOKINGS INSTITUTION'S CAPACITY STUDY

The Brookings Institution attempted to measure America's capacity to produce during the period from 1900 to 1930 as part of a more comprehensive study entitled "The Distribution of Wealth and Income in Relation to Economic Progress."[9] Special attention was devoted to

---

[8] Edwin G. Nourse and Associates, op. cit.; Loeb and Associates, Report of the National Survey of Potential Product Capacity, New York Housing Authority, 1934; and National Resources Committee, Patterns of Resource Use, Washington, 1938.

[9] The other volumes in this series are: M. Leven, H. G. Moulton, and C. Warburton, America's Capacity to Consume, 1934; H. G. Moulton, The Formation of Capital, 1935; H. G. Moulton, Income and Economic Progress, 1935; and E. G. Nourse and H. B. Drury, Industrial Price Policies and Economic Progress, 1938.

production in 1929, and, in appraising the study, we shall be primarily concerned with the measure for this year. Broadly, the authors sought a measure of capacity which would indicate how much output could be produced in a functioning economy if resources were adequately distributed. In the words of the authors, the objectives are as follows: "Our conclusions will be put in terms of the total quantity of consumers' goods which our 1929 plant and personnel were capable of turning out in substantially steady flow;" and "We shall throughout limit our estimates to what would be practicably attainable under conditions of 'sustained simultaneous operation.' "[10]

The authors claim that the measurements of capacity are made on the basis of technological considerations alone and that the problem of costs is ruled out. They interpret excess capacity as being technically unnecessary idleness.

The general procedure for the measurement of capacity was similar to 1) above. The authors divided the economy into three sectors: the raw materials, fabrication, and service sectors; and each of these was divided into industries. They accepted the distribution of the mechanical facilities in the economy and did not consider any additions which might be made during the year or the conversion of facilities from one line to another. Capacity estimates were then computed for every industry where a meaningful capacity measure could be found. These estimates were the maximum output which could be produced on the basis of present techniques in the industry and prevailing customs as to the number of shifts and the length of the working day. Adjustments were made for seasonal variations in output, necessary interruptions caused by such things as maintenance, repair, and breakdown of machinery, and for the capacity of idle plants and machinery. The ratio between actual output and capacity was taken to give an index of the rate at which individual industries were producing. A weighted average of these was computed to give a capacity measure for the entire economy and the rate at which it was producing.

For the year 1929, the weighted ratio of utilization for the economy is 80 percent. According to the authors, 25 percent is then an estimate of how much larger output in 1929 could have been with our present resources and methods of production. To allow for failures of coordination and possible decreases in technical efficiency as output is increased, this estimate is reduced to 19 percent. Since the Brookings Institution was interested in a practical measure of capacity, in a later chapter it considered the problem of the adequacy of the labor force to implement the capacity output. The authors discovered that, if the economy had operated at full capacity in 1929, the manufacturing industries would have needed roughly one million more men than were habitually employed in those industries, and mining industries twenty thousand more men. They claim that this would not have caused a serious bottleneck, since about 1,005,000 men would not have been required even at capacity output in other industries. Thus, with some redistribution, the labor forces would have proved adequate for full utilization of the industrial plant.

In what sense does the measure represent America's capacity to produce?

Let us examine, first, the contention that the measure is dependent only on technological considerations and that prices are of no concern to the authors. It is clear that, though their measure has a nonprice bias, they have not succeeded in determining maximum output conditioned only by technological factors. David Rockefeller and Arthur F. Burns wrote critical

---

[10] Edwin G. Nourse, op. cit., pp. 22 and 23.

analyses of the study and, in essence, made the following comments on this point.[11] The authors were unable to avoid making implicit assumptions about prices since to make estimates of capacity on the basis of prevailing techniques is to assume the whole fabric of prevailing price relationships. The technical methods of production cannot be determined independently of economic forces such as demand and costs, and, if these change, it is to be expected that at least some technical methods will be changed or adjusted. Even if we ignore this consideration, whether an idle plant or an idle machine is obsolete can only be determined by the value of their potential services. In addition, value measures are needed to determine the adjustments to be made for maintenance and repair time.

Apart from these minor failures of the Brookings study, the over-all approach was inadequate to answer the question they posed. As we have indicated previously, the authors set up the following problem: "Let us emphasize that at all stages of our discussion we are concerned not with theoretical maxima of production but with practical results which could be obtained under conditions of operation with which we have had actual experience."[12] To repeat, it appears that they sought a measure to indicate how much output could be produced if resources were adequately distributed and their method, outlined above, indicates that their result cannot be accepted as an adequate answer to the problem they themselves raise.

While a realization of their capacity output would entail some substantial redistribution of resources (labor and materials), it cannot be construed as a measure of the efficient allocation of resources, as in 2), either in a general sense in which all resources must be considered as mobile or in the more limited sense of the study where mechanical facilities are used only for the production of their present commodities and labor has the unique attribute of easy mobility. In the latter sense of a capacity, without a system of relative values, which the study abstracts from in the main, a unique capacity output cannot be determined for the economy. All that can be done, theoretically, is to find a transformation function showing the various maximum alternative combinations of goods which the economy is capable of producing, but there will be no way to choose among these. A choice necessitates a system of values. Assuming, for the moment, that we were interested in maximizing the real value of the national product, then, with some value system, we could select one of these positions such that no additions to the value of national product could be realized by shifting any resource to some other employment.

The authors have not explicitly considered value, so it is impossible to envisage what type of price system they foresaw for production at full capacity. However, there is some evidence to support the argument that they were really interested in a measure of capacity output under a price system and distribution of income different from that prevailing in 1929. An underconsumption, oversaving viewpoint pervades the study. In a later volume the remedy recommended for our economic ills is to increase the purchasing power of the masses by passing on to consumers the benefits of technical progress through reductions in prices. A quote from the first volume is indicative of this outlook. "Evidently, we have not succeeded in utilizing our opportunities fully, and thus poverty has persisted where apparently it might have been reduced, if not removed. This has caused the question to be raised again and again whether

---

[11] David Rockefeller, Unused Resources and Economic Waste, University of Chicago Press, 1941, pp. 192-193; and Arthur F. Burns, loc. cit., pp. 479-480.

[12] Edwin G. Nourse, op. cit., p. 23. In this quotation "theoretical maxima of production" refers to the output which could be attained if no allowances were made for maintenance and repairs of equipment, whereas "practical results" imply the output potential after deductions for these factors.

existing conditions as to the distribution of national income were responsible for maladjustments in the economic system which retarded economic progress."[13] If it was their intent to measure capacity output in a manner similar to 3), again they have failed by not specifying the system of prices and the distribution of income.

A final point which militates against the acceptance of the measure as practicable is that, in effect, it is a timeless analysis. The authors derive a final position for the economy after all adjustments have occurred without an analysis of the adjustments themselves. To achieve their capacity output, some of the necessary adjustments, e.g., involving labor, could probably be accomplished only under some central direction and would take a number of years at substantial cost. In addition, the authors have dismissed the labor problem quite summarily by merely determining whether there existed a sufficient number of workers to fill the necessary number of positions. They do not even discuss the matter of difference in training and skills of the labor force.[14]

We must conclude that after reading this study one would have little idea as to what the capacity output of the United States was in 1929 and that, consequently, the volume was a failure. The study does point up the important fact that in dealing with the notion of a capacity one must proceed with extreme care.

In our brief analysis of the first approach we have tried to indicate the indefinite nature of a capacity concept and the necessity, for any particular problem, of making a careful selection from among the many feasible measures of a capacity. In Part II we discuss the second general approach to the problem.

Part II
OPTIMUM STOCK LEVELS

THE BALANCING OF COSTS APPROACH

The second approach to the study of capacity is useful for analyses of investment behavior. Again, a serious problem is that the concept is frequently used ambiguously. For example, a crucial deficiency of the acceleration principle model is that it does not allow for excess capacity, but, even in critical studies of the principle, it is not usual to find a clear-cut discussion of excess capacity, or even capacity, for that matter.[15] Before analyzing investment models, which is the main subject of Part II, we shall discuss what the balancing of costs entails and how a capacity concept may be derived.

Consider an economic enterprise attempting to maximize profits from the sale of its commodity. A technological production relation connects quantities of raw materials, labor, and services of durable capital goods with quantities of the commodity. A problem confronting the enterprise is to determine the stocks of the inputs which it should hold at any time. Periodically, it has to decide what should be the appropriate size of the plant, the quantities and types of machinery, and the average inventory levels of raw materials and finished goods. In principle, the problem is solved by balancing estimated costs during some planning period such that for estimated demands and prices in this period total costs are a minimum.

---

[13] Edwin G. Nourse, op. cit., p. 6.

[14] For a discussion of their labor estimates cf. Arthur F. Burns, loc. cit., pp. 487-489, and David Rockefeller, op. cit., pp. 194-195.

[15] An exception is an article by A. D. Knox, "The Acceleration Principle and the Theory of Investment; a Survey," Economica, August 1952, pp. 276-279.

The solution determines the optimum levels of the stocks of goods which the firm should hold, and these levels of stocks jointly may be called the capacity for the firm.[16] This capacity tells us what is the optimum physical size of the enterprise, and when the physical structure is too large or too small excess or under-capacity exists and adjustments in stock levels are profitable. These adjustments in stocks are the essence of the problem of investment.

In practice the balancing of costs technique is difficult and instead firms may use rules of thumb or the entrepreneur's "feel" for the market to determine investment plans.[17] Whatever method is actually used by the entrepreneur, the main task for the prediction of investment is to estimate the desired levels of stocks. Then a comparison with the existing levels of stocks and a knowledge of the process of investment, i.e., the realization over time of investment plans, would permit estimates to be made of the rate of investment.

In the remaining sections we shall examine present day models in economics which treat in various ways the problem of investment in durable goods.

THE ACCELERATION PRINCIPLE

In discussions of investment it is customary to distinguish between net investment which is the algebraic sum of net changes in stocks of durable production goods and replacement investment which is made to replace obsolete stocks, e.g., replacement of an old metal lathe. Replacement investment cannot be less than zero, whereas net investment may be positive, zero, or negative. Some investment models, in particular the acceleration principle model, assume that the categories are distinguishable. In principle this seems justified as the motive for replacement investment is to maintain an existing organization while net investment is made to change the size of the organization. In practice, however, even replacement may change the size of the organization in the sense that it adds to the income stream if the unit replaces obsolete machinery with technically superior equipment. F. A. Hayek has argued that to distinguish an act of investment as net or replacement it must be determined "whether a person maintains a stock of nonpermanent resources which will secure him an increasing, constant, or decreasing income stream, not whether the stock itself increases, remains constant, or decreases in any of its directly measurable dimensions."[18]

The timing of replacement investment also presents uncomfortable problems as capital equipment is usually not of the one-horse shay variety but has a maturity dependent upon the intensity of use of the equipment and the rate of technological progress. Investment models which are concerned with predicting net investment are able to retain their simplicity only by assuming away the complications caused by replacement investment. These models generally

---

16 Usually when capacity in this context is considered, authors mean only the desired quantities of durable goods, such as the plant and the machinery.

17 Important work has recently been done on the determination of optimum inventory levels and is attracting increasing attention from business and military organizations. Cf. Thomson M. Whitin, The Theory of Inventory Management, Princeton: Princeton University Press, 1953; Kenneth J. Arrow, Theodore Harris, and Jacob Marschak: "Optimal Inventory Policy," Econometrica, Vol. 19, No. 3, July 1951, pp. 250-272; A. Dvoretzky, J. Kiefer, and J. Wolfowitz: "The Inventory Problem: I. Case of Known Distribution of Demand," Econometrica, Vol. 20, No. 2, April 1952, pp. 187-222; and J. Laderman, S. B. Littauer, and Lionel Weiss, "The Inventory Problem," Journal of the American Statistical Association, Vol. 48, No. 264, Dec. 1953, pp. 717-732.

18 F. A. Hayek, The Pure Theory of Capital, London, 1941, pp. 300-301. For more complete discussions of net and replacement investment cf. A. D. Knox, loc. cit., pp. 281-290, and Simon Kuznets, "Relation between Capital Goods and Finished Products in the Business Cycle" in Economic Essays in Honour of Wesley Claire Mitchell, New York: Columbia University Press, 1935, pp. 209-267.

handle this category by assuming that durable goods have a fixed maturity and by abstracting from technological change.

It is well to begin our discussion with the acceleration principle as most investment models are variations of this form. The empirical evidence against the principle is sufficient to cause us to lose interest in this model itself, but it is worthwhile to discuss why the tests have been negative and to see how other models have developed from its deficiencies.[19]

The acceleration principle model attempts to determine the amount of net investment over time for some economic unit.[20] The unit is usually considered to be a firm, an industry, or the economy, though little attempt has been made to distinguish conceptually among the models for these different levels. The principle states that net investment for an economic unit, e.g., the firm, depends on the rate of change of its real output. The basic relation is $I_{t+\alpha} = \beta (Y_t - Y_{t-1})$, where $I_{t+\alpha}$ is net investment in period $t+\alpha$, β is some constant, and Y_t is output in period t. In general, the time lags must be determined empirically. The basis for the principle has been stated to be approximately the following: there is a technological relation between stocks of durable goods necessary for production and output such that as output varies it becomes necessary to change the level of these stocks or, in other words, the demand for finished goods is reflected in the demand for capital goods.[21] This statement, however, does not bring us to the equation above. To derive this equation, we may use any of several methods concerning interpretations of the constant, β, where different interpretations yield essentially different relations.

One path is to interpret β as expressing a production relation between services from capital goods and output.[22] With this interpretation it can be shown that a constant β, implies either constant returns to scale, assuming fixed factor price ratios, or, if in addition to constant returns to scale we assume fixed coefficients of production, no assumptions about price ratios are necessary. However, we are still a step removed from the stock-output relation postulated in the principle since a constant ratio between services from capital goods and output does not necessarily imply a fixed stock-output ratio. To arrive at the equation for investment we have to assume that durable goods have a fixed maturity and that their intensity of use remains constant.

The interpretation of the principle as essentially a production function deprives it of any general validity due to the restricted shape of the function it postulates and strongly suggests that in this form it would not be very useful for predicting investment.[23] Moreover, for

---

[19] For empirical tests of the acceleration principle cf. Simon Kuznets, op. cit., pp. 248-267; J. Tinbergen, "Statistical Evidence on the Acceleration Principle," Economica, 1938, and Statistical Testing of Business Cycle Theories, League of Nations, Geneva, 1938, Vol. I, Chapters 3 and 5 and Vol. II, Chapter 2; and Hollis B. Chenery, "Overcapacity and the Acceleration Principle," Econometrica, Vol. 20, No. 1, January 1952, pp. 18-28.

[20] The acceleration principle has also been used to determine investment in inventories of raw materials, goods-in-process, and finished goods. However, empirical work by Moses Abramovitz, The Role of Inventories in Business Cycles, 1948, and Inventories and Business Cycles, New York: National Bureau of Economic Research, 1950, and the analytical studies quoted above have shown that, as an explanation of inventory investment, the acceleration principle is clearly inadequate.

[21] J. M. Clark, "Business Acceleration and the Law of Demand," Journal of Political Economy, March 1917, p. 217.

[22] Cf. R. S. Eckaus, "The Acceleration Principle Reconsidered," Quarterly Journal of Economics, Vol. LXVII, No. 2, May 1953, pp. 211-216. In his discussion, Eckaus considers a production relation between capital stock and output, whereas it is customary to set up a function between services and output. A constant relation between capital stock and output requires assumptions which Eckaus does not make explicit.

[23] Authors testing the principle statistically frequently use this or a similar interpretation; cf. Simon Kuznets, op. cit., pp. 248-267.

a profit maximizing firm it is only under certain conditions that it would be plausible to predict its investment from its production function. Among the conditions are these: excess capacity does not exist at any point of time; sufficient funds are available to implement any investment plans; there are no physical restrictions on net investment, either positive or negative; there are no indivisibilities of capital equipment; the nature of expectations is such that when there is a change in output the belief is that output will remain at a level sufficient to give an adequate return on money outlays. The assumptions necessary to make this interpretation of the principle plausible are so restrictive and so much must be explained outside the model that it seems reasonable to discard this form of the relation.

Other interpretations are conceivable. We could regard the coefficient, β, as a capacity index, though not necessarily identical to the one developed previously. A. D. Knox, for example, has outlined the possible definitions of capacity for an interpretation of β. He accepts output at the minimum point of the average total cost curve as the most useful and, with this definition, offers the following rationale for the principle. "The acceleration principle may be given a simple rational basis. It may be interpreted to mean that entrepreneurs will meet a rise in demand by expanding their plant, where the cost of producing the extra output with the existing plant exceeds the operating cost with the enlarged plant plus the costs of purchasing and installing it. If we assume with the acceleration principle that changes in output are the only forces making for investment, we may conclude that there is no incentive for investment before the least cost point is reached, and an increasing incentive the further beyond that point output goes."[24] With this interpretation the principle loses its simplicity and again the basic form must be abandoned. Knox thinks this viewpoint of the model is particularly damaging to timing aspects of investment, while it may still be useful in determining the volume.[25]

Economists, of course, have not regarded the basic form as an unqualifiedly acceptable description of empirical investment behavior. A persistent criticism is that the acceleration principle is active only in phases since, if excess capacity exists, the model is inoperative. Disregarding other limitations for the monent, this means that if the firm is at capacity the acceleration model is sufficient to explain investment behavior when the output rate changes, and if the firm has excess capacity, the model cannot predict investment until the excess capacity is worked off. In this criticism capacity seems to refer to the desired quantity of capital stock which, as we have discussed, is determined from both cost and revenue functions. The desired level of capital stock is equivalent to the least cost point, as in Knox's interpretation, only under special conditions. From criticisms of this type other investment models have developed. In the next section we shall consider variations of the acceleration principle where desired levels of stocks enter explicitly into the investment function.

OTHER INVESTMENT MODELS

Apart from models which introduce desired levels of stocks, some variations are concerned solely with the timing of investment and the time dependence of investment on past output. Since these latter models do not make any fundamental changes in the basic form, we shall not discuss them further.[26] A. D. Knox, in his conclusions on the acceleration principle,

---

[24] A. D. Knox, loc. cit., p. 279. For the discussion of capacity, see pp. 276-279.
[25] Ibid., pp. 279-281.
[26] For such a model cf. J. R. Hicks, A Contribution to the Theory of the Trade Cycle, Oxford, 1950. Generally, Hicks makes investment in period t a weighted function of an indefinite number of past changes in output.

sums up the determinants of investment in the following form: $I = \beta(K_D - K_A)$, where β is the inverse of the gestation period, I refers to the rate of gross investment per unit of time and not to the decision to invest, K_D is the desired level of capital stock, and K_A is the part of existing equipment efficient enough to be retained in use. The rate of investment is thus made to depend upon differences between the actual level and the desired level of capital stock. Knox correctly emphasizes that this is a summary behind which lie some very complex relationships.[27] In particular, these complex relationships relate to the determination of K_D, which, as we have discussed previously, is the crux of the investment problem. In addition, β, itself, may not be simply determined. In other words, it is necessary to solve the problem of investment for the firm before we can set up Knox's summary function. Other variations are akin to Knox's and are, similarly, more in the nature of summary relations than theories of investment.

Richard M. Goodwin derives a relation which in essence is identical to Knox's except that it relates to net investment.[28] As does Knox's, Goodwin's function involves closing the gap between the desired levels of capital stock and the actual level during a gestation period. The relation, as follows, is given in differential form and is for the economy as a whole: $\dot{K} = \frac{1}{\xi}(kY + \Phi - K)$, where K is the actual quantity of capital, $\dot{K}$ is the first derivative of K with respect to time or the rate of investment, ξ is the period of time necessary to close the gap, k is the acceleration coefficient, Y is income, and Φ is an innovational shift function which allows for variations in interest rates and changes in expectations. $kY + \Phi$ may be considered as the desired level of capital stock. When this is equal to the actual level of capital stock an equilibrium position is reached, and there is no net investment. $kY + \Phi$ is thus the equilibrium level of capital which is reached in ξ years if nothing else changes in that span of time. Goodwin considers this to be a flexible form of the accelerator which allows for the major shortcoming of the ordinary accelerator. He says, "The ordinary accelerator implies that capital is always perfectly adjusted to income. This is its major shortcoming, for we know that there are times of excess capacity, and other times of short capacity."[29] Again, Goodwin's formula is a summary relation which would require further analysis before it could be tested empirically.

A similar but more interesting statement of an investment relation is derived by Hollis B. Chenery in which the author attempts to determine analytically the desired levels of capital stock under certain conditions.[30] It should be noted that Chenery defines capacity as the point of tangency of the plant curve, which is a cost curve for a given plant, and the long run cost curve.[31] The optimum stock level in his scheme is what he calls the optimum degree of overcapacity and he determines this optimum stock level explicitly under the following conditions. 1) The entrepreneur is profit-maximizing. 2) There is an expected secular rise in demand. 3) The price of the product and the production function for the firm are given. The problem is to find the optimum size of the firm if economies of scale exist and the general result is the following. "It is shown that when economies of scale exist, excess capacity will occur even with perfect forecasting: this may be called 'optimum' overcapacity. The degree of

---

[27] A. D. Knox, loc. cit., pp. 295-297.
[28] Richard M. Goodwin, "Secular and Cyclical Aspects of the Multiplier and the Accelerator," in Income, Employment, and Public Policy, Essays in Honor of Alvin H. Hansen, W. W. Norton and Co., Inc., New York, 1948, pp. 108-132.
[29] Richard M. Goodwin, op. cit., p. 120
[30] Hollis B. Chenery, loc. cit., pp. 1-28.
[31] Ibid., p. 14, footnote 25.

overcapacity is expressed as a function of the properties of the production function, the planning period, and the discount rate."[32] Different solutions are indicated for various given values of the planning period, n, the discount rate, r, and the scale coefficient, ψ, which is the elasticity of output with respect to cost. Chenery does not explicitly compute optimum overcapacity for fluctuating output but indicates that the consequent changes in the parameters will merely reduce the optimum degree of overcapacity.[33]

According to Chenery, when the production function is known and the parameters r and n can be estimated, λ, the degree of utilization of plant can be determined. Then, if we begin with a period of equilibrium when capital and output have their "normal" relationship and if we select a suitable lag,[34] we can write $\lambda K_t = \beta X_{t-\theta}$ as the relation between capital and output where λ is as defined, K_t is the optimum level of capital in this period, β is the accelerator constant, X is output, and θ is the lag between output and adjustments in capital stock. The acceleration principle is then reformulated so that investment is assumed to be proportional to the deviation of actual output from "normal" output. The resulting formula is called the capacity principle: $\Delta K_{t+\theta} = b(\beta X_t - \lambda K_t)$, where b is a reaction coefficient, similar to Goodwin's.[35] The value of λ does not vary with changes in demand unless the scale coefficient varies with plant size, but the absolute size of plants which it pays to build will vary with such changes.

Chenery made comparative statistical tests of his hypothesis against the simple acceleration principle hypothesis in six basic industries. This was a test in which relevant parameters in each case were determined by simple correlation and regression analysis. The result, according to Chenery, was that the capacity principle showed reasonable success for all industries while the accelerator was inadequate in half. In two cases, though, the acceleration principle was slightly better than the capacity hypothesis.[36]

There are other investment models in economics but we shall not discuss them here. These tend to be similar, except for emphasis, to the ones mentioned above. As Eckaus has noted, "Although the relations were generally derived by different methods there is a strong 'family resemblance' among the investment-determining relations developed by Tinbergen, Klein, Kalecki, Kaldor, Chenery and Goodwin."[37]

SUMMARY

We have analyzed two approaches to the study of capacity. The first involves finding maximum output where the solution is determinate only if the conditions under which the output is produced are completely specified. The appropriate set of conditions depends on the

---

[32] Hollis B. Chenery, loc. cit., p. 2.

[33] Ibid., p. 11.

[34] "Normal" output is not clearly defined. In a footnote on page 13, Chenery says only that "The 'normal' output represents the capacity of the industry (under stationary conditions) in the sense of the most profitable output, but not in the sense of maximum output with given plant."

[35] Presumably, the derivation of this formula is as follows:

a) $\lambda K_t = \beta X_{t-\theta}$ b) $X_{t-\theta} = \dfrac{\lambda K_t}{\beta}$

c) $\Delta K_{t+\theta} = b\beta \left(X_t - \dfrac{\lambda K_t}{\beta}\right)$. d) $\Delta K_{t+\theta} = b(\beta X_t - \lambda K_t)$.

Chenery only gives an explicit derivation for a similar formula.

[36] Hollis B. Chenery, loc. cit., pp. 18-28.

[37] R. S. Eckaus, loc. cit., p. 233. References to models by Tinbergen, Klein, Kalecki, and Kaldor may be found in Eckaus's paper.

particular problem at hand, and, consequently, numerous measures of capacity are conceivable for the same commodity or combinations of commodities. For this approach we explicitly examined three common measures of capacity. One is usually called technological capacity and relates to the productive capabilities of plants and machinery without regard to the costs of complementary materials and the demand for the product. A number of American industries attempt to compute capacity on this basis. The other two, which consider alternative uses of resources among the specified conditions, may be called economic measures of capacity. Little work has been done in this area though economic measures of capacity have important uses. For example, the problem of finding a system of priorities and allocations for a defense system or war economy is dependent on measures of this type. Despite their deficiencies, input-output studies which attempt to determine the capabilities of the economy to produce various selected bills of goods may be considered as a step in the direction of measuring economic capacity.

The second approach concerns optimum stock levels and is useful for studies of investment. The principal explanation of investment behavior, the acceleration principle, and some variants of this form were discussed. The acceleration principle explanation has proved to be a failure and a major reason is that it ignores the effects of existing capital goods on investment. Most variants which attempt to correct for this deficiency are not completely satisfactory since an analysis of investment plans is frequently eschewed for a summary relationship in which the desired level of capital stock is a given parameter. One exception is Chenery's model in which the author attempts to determine analytically the optimum level of stocks under special conditions. Yet, despite the improvements made in this model, the acceleration principle yielded better results in two out of six statistical tests.

While it may be wishful thinking to believe that such a volatile activity as investment can be predicted accurately by models, considerable room for improvement remains. It is difficult to suggest what direction research should take; perhaps less attention should be devoted to finding a unique model for predicting investment and more to examining the peculiarities of particular industries. A first step is certainly a careful analysis of how investment plans are made for both net and replacement investment.

* * *

ISSUE PRIORITY: LAST IN FIRST OUT (LIFO) VS FIRST IN FIRST OUT (FIFO) AS A METHOD OF ISSUING ITEMS FROM SUPPLY STORAGE

Joseph Albert Greenwood
Quality Control Division,
Bureau of Aeronautics, Navy Department

Items may be issued from storage by the principle of "first in, first out" or "last in, first out." This paper gives comparisons of the two methods by various criteria.

INTRODUCTION

The objective of this paper is to make a comparative analysis of two opposing systems of issuing shelf-deteriorable items from storage. These are (1) FIFO (first in first out) in which the oldest material available is issued, and (2) LIFO (last in first out) in which the newest material available is issued. FIFO appears to be the method more widely used now, but under the assumptions of this paper the results indicate a likely wide field of application in which LIFO would be the more profitable method to use. Formulas are given for different cases from which comparison of the two methods can be made.

A strong assumption of steady demand from the field is made in order to simplify obtaining explicit formulas for the various conparative properties, such as average age of items in the field. The implications of this paper should certainly be of interest to the military since in many cases large war reserves must be maintained in storage, and the results apply most strongly to such situations. The more practical and difficult approach of introducing stochastic variation into actual shelf life, actual field life, and thence to the demand from the field remains for later consideration. This would be especially desirable when the inventory is so low that the risk of failing to meet the demand might be appreciable due to the demand variation. Except for the latter contingency, it is believed that the results obtained herein for steady demand or perfect predictability are sufficiently valid to warrant decision and action based on them. The main reason for presenting this introductory paper on the subject is to demonstrate the existence of this important and practical field which has not apparently otherwise been investigated, in the hope of stimulating more thorough research along these lines.

Before carrying the discussion further, some of the basic terms will be clarified. Assigned shelf life is an arbitrary time that a new item may be left on the shelf and still be issued to the field. If unissued by that time, it is over age and is either discarded or reworked. True storage age is the length of time a new item can sit on the shelf before deteriorating to a preset level of utility or quality. This level is such that an item of its identical quality in the field would be discarded, and in fact represents the highest quality level at which this would still be true. The field life is the length of time a new item would be useful in the field. Input is the number of new items placed into the system each unit of time.

The nearer the assigned shelf life is to the true shelf life, the more nearly does the FIFO system tend to prevent any material from going overage on the shelf, regardless of the

rate of input of new items to storage. This sometimes fosters the illusion that excessive numbers are required as input in order to maintain given field requirements, and as a kind of corollary, it seems to be considered by some that if a given system will only prevent over age items on the shelf, then no extra cost for storage is involved and all must be well.

A little reflection shows that there is a steady penalty payment going on for the privilege of maintaining items in storage, since each item is steadily losing some of its future field usefulness. By changing the method of issue to LIFO or last in first out, i.e., issuing the newest material available from the shelf, this necessary payment in terms of potential life units is shifted from the best to the worst. Since these best life units that were formerly used up as shelf "payment" are now issued to the field, they represent fewer malfunctions in use, perhaps better operation, and the latest incorporated changes. The items will also last longer in the field, thus saving on what may well be expensive change-overs in the field. Under the FIFO system of issue there may be a general unawareness that the actual field life is far shorter than would be the case with newer material that would be issued under LIFO. In other words the FIFO system makes it difficult to obtain the experience required to know what should be expected under a possibly more adequate system of issue. The FIFO system takes those best life units on the shelf as payment for storage. The LIFO system puts as many as possible of those best life units directly into the field and forces the system to take an equal number of the worst life units as payment.

While the general comparative philosophy is stated above, in actuality if the assigned storage life is set arbitrarily less than the true storage life, there is a further payment exacted from the LIFO system in terms of these worst life units, for equal numbers maintained both in storage and in the field in comparing the two systems. Therein enters the consideration of the savings due to decreased rate of turnover in the field and the aforementioned intangible benefits possible under LIFO, as compared with the extra amount of new input required to maintain the system under LIFO. Under rather general conditions, formulas are obtained which contrast the two systems in several important respects, such as average age of items in the field, and which allow one to decide in any particular case which system would be the more profitable, financially. These are found for the case of a straight line deterioration curve, and the results are more strongly in favor of LIFO if it is known that the deterioration curve of remaining field life as a function of elapsed storage time is concave up (the solid curve of Figure 1).

The curves drawn in Figure 1 denote a class of possible deterioration curves. Discussions with engineers working with or responsible for the specific items leads the author to believe that most practical situations will fall in this class. However, it is conceivable that the curve might level off for a ways or have both convex and concave features. These possibilities also remain to be studied.

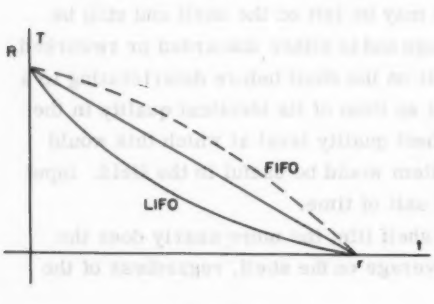


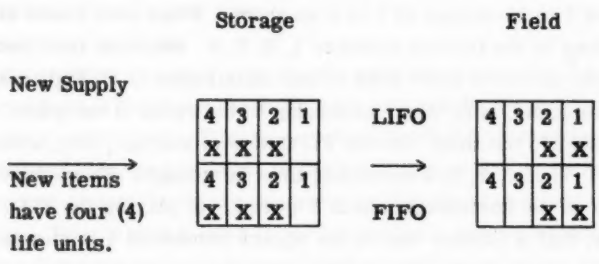
Figure 1

Not formally discussed in this paper is the possibility of further decreasing the costs of the LIFO system by salvage of or finding less critical use for the items that do go over age on the shelf. Thus an overseas activity might be strictly on the LIFO end of issue and an activity closer to the point of issue might be able to use profitably those that were on the lower end of shelf life, if not actually over age. In this

connection, it would seem that the optimum utilization of the total life units involved would be to let the second user reach into the stock as far up the scale toward better material as possible, consistent with being reasonably certain that the prime LIFO user would never have to dip down below the actual place in terms of storage age that the secondary user had been permitted to receive.

A good feature of the LIFO system is that the assigned storage life may be set larger than would be practicable under FIFO, because the oldest items would not be issued anyway except in emergency when they would likely not need a long life. This longer assigned storage life would formally introduce still further decrease in cost of the LIFO system, for the purpose of making the actual comparison and deciding on which system to use. It should not be overlooked that it is inherent in the LIFO system that items may go over age on the shelf in storage. However, in a real sense these items represent the "payment" mentioned earlier which exists but is partially if not entirely concealed in the FIFO system.

Before going into the actual analysis of the systems, the comparative features will be illustrated by a "game" described below. In order to keep it simple it is assumed that a new item is available whenever needed. Thus under LIFO no item in storage is ever sent to the field. However, even this artificial setup takes on an aspect of reality when viewed in the light of a relatively large storage inventory. The results in this paper include also the case of periodic input, for example buying new stocks every six months to replenish the system, but to play the "game" corresponding to the more realistic and complicated situations requires considerable time for it to settle down and average out to the results obtained in this paper. The straight line or linear deterioration curve and storage life equal to field life of four time units are assumed for this example. The comparative game situation is depicted below.



An X denotes an item having the number of remaining units of life indicated by the associated number in the square with it. It is assumed that two items are maintained in the field and three in storage at all times, and that a new item is available whenever needed. To "play the game," assume that one unit of time has passed and immediately move every item in both field and storage one square to the right, since they have each lost one life unit. Replenishments are made according to the rule of each system. In FIFO the right hand item in storage is set over into the field, when needed, at the same number position as the square at which it rested in storage. New items from the new supply are placed in square 4 in storage. Under LIFO, when an item moves from number 1 in the field on out of the rectangle to the right (meaning it is then to be discarded or reworked), a new item is placed at square 4 in the field directly from the new supply. The result of this play is a repetitive cycle of four moves from which it is easy to compute that the average remaining life of items sent to the field under FIFO is 1.6 life units and under LIFO is 4 life units. Thus on the average, under FIFO

each item that reaches the field has spent 60 percent of its useful life in storage while under LIFO none of its useful life was spent in storage. The rate of replacement in the field for FIFO is 2-1/2 times that for LIFO. Note that the same rate of input of new items was required for both systems.

The LIFO system has a natural "therapeutic" effect on a system made "ill" by a FIFO method of issue. It has already been pointed out that under FIFO a state of affairs may exist in which the turnover in the field is quite rapid with little if any going over age in storage. The input of new items may be quite high to maintain such a system. Such a state of affairs tends to get worse under FIFO. Introduction of the LIFO system immediately calls a halt to that, and large numbers of items begin going over age in storage. When stability is reached, large numbers may still be going over age in storage. It is then clear to all that such a price must be paid only to maintain the large number in storage, not primarily to maintain the numbers in the field. A proper decision can then be made as to the amount of future procurement required, and in fact, the formulas are at hand to tell one what the input should be to maintain stated numbers in field and storage.

If it is desired to illustrate the above feature by means of the "game," use the same setup as before, but have only two items in storage, at positions 3 and 4. Then at the end of each time unit, introduce two new items into each system. The LIFO system will approximate a ratio of seven over age in storage to three over age in the field, while the FIFO system will allow none to go over age in storage. Yet the same number was maintained in the field and the same input of new items was made for both systems.

To illustrate the fact that under a concave-up deterioration curve (Figure 1) LIFO requires a smaller input of new items to maintain fixed numbers in the field and in storage than does FIFO, use a setup which has the life span divided into 10 equal parts. The squares will thus be numbered 1 to 10 instead of 1 to 4 as above. Place four items in storage at numbers 7, 8, 9, 10 and four in the field at numbers 1, 2, 3, 4. Maintain four items in both storage and field. The concave-up curve to be used in this illustration is in Figure 2. It will be applied to the FIFO side of the game, only, since according to the rules of the game, no items in storage under LIFO are sent to the field. On the FIFO side of storage, only, renumber the squares from right to left 9, 8, 7, . . ., 2, 1, 0 indicating time in storage. Those numbers also correspond to the numbers on the horizontal axis of Figure 2. In playing the FIFO side of the game, suppose, for example, that a counter was in the square numbered 4 in storage and this counter

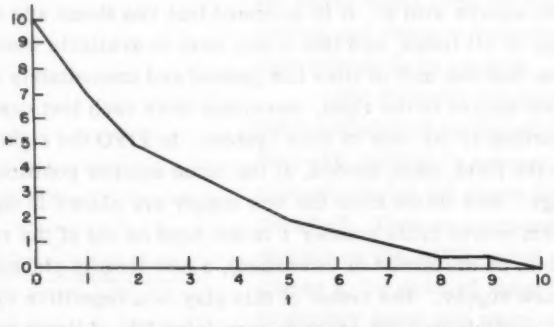


Figure 2 - Remaining field life as a function of elapsed storage life

is due to be moved immediately to the field. According to Figure 2, a 4 on the horizontal axis corresponds to a 3 on the vertical axis by virtue of the curve, and so this counter would be placed in the field at number 3.

Since the systems have cycles of 7 and 10 moves, respectively, it is easily found that the following tabulation gives the results of 70 moves of each system.

	LIFO	FIFO
Total input of new items	56	80
Over age in storage	28	0
Over age in field	28	80

Thus the input of new items required under FIFO is 43 percent greater than for LIFO.

GENERAL ANALYSIS

In this paper the problem is attacked on an "average" basis, only, although it is clearly a situation in which probability distribution theory should play a part (it has been suggested also that renewal theory might apply) to give more definitive results and enlarge the scope of application.

It is assumed that one is dealing with a large volume operation in the theory of this approach and that there is a steady demand from the field. Further assume that the curve of remaining field life as a decreasing function of time in storage is a straight line, or concave up, or concave down, as shown in Figure 1. In this figure

t = elapsed storage time of the item

T = remaining field life of the item

r = true storage life of a new item

R = field life of a new item.

Consider the straight line of Figure 1. Its equation is $T = R - \frac{R}{r}t$. Under this assumption the proportional part of the field life remaining is equal to the proportional part of the storage life remaining in the item. It is then an important and interesting fact that for any r and R, and for assigned storage life equal to the true storage life r, both LIFO and FIFO require the same input of new items to maintain equal numbers of items in the field and equal numbers in storage. This is easily seen if one attributes a certain number of "life" units to the item. Then this fixed number of life units may be used up in either storage or field, but at possibly different rates per unit time. If fixed numbers of items are maintained in storage and in the field, then regardless of the ages of the items in both places there is a constant loss of life units. Thus all systems of issue, LIFO and FIFO in particular, require the same input of new items to make up for this constant loss. On the other hand, if the deterioration curve is of the concave up type (the solid curve of Figure 1), the tendency is for the LIFO system to require smaller input of new items than the FIFO system to maintain these fixed numbers in storage and field. This stems from the fact that under LIFO the average storage age is greater than under FIFO, i.e., in Figure 1 average t for LIFO is greater than for FIFO. But the rate of change (loss) of life units on the vertical or field scale is less the farther to the right (i.e., the greater) is t. Therefore, at least for the average t value in storage there is less loss of life units under LIFO and, therefore, more total life available for the field for equal inputs to the two systems. Since life units are equivalent to time units in field usage, the two systems have the same field life unit rate usage. Therefore, smaller input is required

for LIFO. For a concave down curve (the broken curve of Figure 1), the reverse tendency is true and smaller input is required for FIFO. The remainder of this paper will be concerned with the straight line situation, only.

The complete set of symbols required for the following results including the derivations of the appendices follows.

r = true storage life of a new item
r' = assigned storage life ($\leq r$)
R = field life of a new item
S = average number of items maintained in storage
F = number of items maintained in field
θ = input of new items per unit time period
A = cost of change-over in the field, i.e., cost of removal and installation, loss of aircraft availability, etc.
B = initial cost of a new item
β = rate of over age in field per unit time
$a = \theta - \beta$ = rate of over age in storage per unit time
$[n]$ denotes "the greatest integer in n." E.g., $[3.8] = 3$, $[4] = 4$.
$\epsilon = r' - [r']$

Case I: Assigned storage life is $r' \leq r$ and a new item is available when needed as a replacement.

The special situation of the above when $r = r'$ has already been partially discussed when it was observed that equal inputs were required for both LIFO and FIFO. It was also the "game" situation discussed.

In the general situation implied by Case I, suppose F, θ, r, r', R, A, and B are given. Then in tabular form seven comparative facts of the two systems will be displayed, with the proofs being reserved for Appendix 1. Immediately following that is given an inequality requiring minimum information which tells whether it would be financially profitable to use the LIFO system instead of the FIFO system.

If it is desired to compare the two systems when stated numbers are maintained in storage and in the field, it would be more convenient to have these seven tabulated facts in terms of S rather than θ. Then the required input could be easily calculated. This also is given in similar form and was accomplished mainly by solving the previous results for S, assuming that θ was given.

F, θ, r, r', R, A, B given.

	LIFO	FIFO
(i) average storage age of items at time they are sent to field	0	$r\left(1 - \frac{F}{\theta R}\right)$ if $< r'$, otherwise r'
(ii) number of items over age in field per unit time	$\frac{F}{R}$	θ if (i) $< r'$, otherwise $\frac{Fr}{R(r - r')}$ if (i) $= r'$.

	LIFO	FIFO
(iii) average age of field items (in field life units)	$\frac{R}{2}$	$\frac{R}{2}(2 - \frac{F}{\theta R})$ if (i) $< r'$, $\frac{R}{2}(1 + \frac{r'}{r})$ if (i) $= r'$.
(iv) input of new items per unit time	θ	θ
(v) average number of items maintained in storage	$S = r'(\theta - \frac{F}{R})$	$S = \frac{r}{R}(\theta R - F)$ if (i) $< r'$ $= \theta r'$ if (i) $= r'$.
(vi) rate of over age in storage per unit time	$\theta - \frac{F}{R}$	0 if (i) $< r'$ $\theta - \frac{Fr}{R(r - r')}$ if (i) $= r'$.
(vii) average age of stored items	$\frac{r'}{2}$	$\frac{r}{2}(1 - \frac{F}{\theta R})$ if (i) $< r'$ $\frac{r'}{2}$ if (i) $= r'$.

Now and then reference will be made to certain of the preceding characteristics merely by quoting the designator (i), (ii), etc.

With the same input θ to the two systems, a smaller number of items is maintained in storage under LIFO than under FIFO. This is a direct result of the fractional waste of life of those items that go over age on the shelf under LIFO. It is thus fair perhaps to compare the two systems as follows. If the cost of the excess change-overs in the field under FIFO exceeds the cost of the wasted life units on the shelf under LIFO, one would save money by using the LIFO system.

If (i) $< r'$ under FIFO, this comparison is given immediately by the equation

$$(\theta - \frac{F}{R})A \geq (\theta - \frac{F}{R})(\frac{r - r'}{r})B$$

or

$$\frac{A}{B} \geq (1 - \frac{r'}{r}).$$

The amount of gain per unit time by using LIFO is then

$$(\theta - \frac{F}{R})\left\{A - (1 - \frac{r'}{r})B\right\}.$$

If (i) $= r'$ under FIFO, the corresponding comparison is given by the equation

$$\left\{\frac{Fr}{R(r - r')} - \frac{F}{R}\right\}A \geq \left\{\theta - \frac{F}{R} - \theta + \frac{Fr}{R(r - r')}\right\}\frac{(r - r')}{r}B$$

or again

$$\frac{A}{B} \geq (1 - \frac{r'}{r}).$$

The amount of gain per unit time by using LIFO is then

$$\frac{Fr'}{R(r - r')} \left\{A - (1 - \frac{r'}{r})B\right\} .$$

F, S, r, r', R, A, B given.

	LIFO	FIFO
(i) average storage age of items at time they are sent to field.	0	$\frac{rRS}{RS + rF}$ if $< r'$ otherwise r'
(ii) number of items over age in field per unit time	$\frac{F}{R}$	$\frac{S}{r} + \frac{F}{R}$ if (i) $< r'$ $\frac{Fr}{R(r - r')}$ if (i) $= r'$.
(iii) average age of field items (in field life units)	$\frac{R}{2}$	$\frac{R}{2}\left(\frac{2RS + rF}{RS + rF}\right)$ if (i) $< r'$ $\frac{R}{2}(1 + \frac{r'}{r})$ if (i) $= r'$.
(iv) input of new items per unit time	$\theta = \frac{S}{r'} + \frac{F}{R}$	$\theta = \frac{S}{r} + \frac{F}{R}$ if (i) $< r'$ $= \frac{S}{r'}$ if (i) $= r'$.
(v) average number of items maintained in storage	S	S
(vi) rate of over age in storage per unit time	$\frac{S}{r'}$	0 if (i) $< r'$ $\frac{S}{r'} - \frac{Fr}{R(r - r')}$ if (i) $= r'$.
(vii) average age of stored items	$\frac{r'}{2}$	$\frac{1}{2}\left(\frac{rRS}{RS + rF}\right)$ if (i) $< r'$ $\frac{r'}{2}$ if (i) $= r'$.

Suppose (i) $< r'$ under FIFO. Then as before, LIFO is indicated if

$$(\frac{S}{r} + \frac{F}{R} - \frac{F}{R})A \geq \frac{S}{r'} B \frac{(r - r')}{r} \quad \text{or} \quad \frac{A}{B} \geq \frac{r}{r'} - 1 .$$

The amount of gain per unit time by using LIFO is then

$$\frac{S}{r}\left\{A - \left(\frac{r}{r'} - 1\right)B\right\}.$$

If (i) = r' under FIFO, LIFO is again indicated if $\frac{A}{B} \geq \frac{r}{r'} - 1$. This is the simplification of

$$\left\{\frac{Fr}{R(r - r')} - \frac{F}{R}\right\}A \geq \left\{\frac{S}{r'} - \left(\frac{S}{r'} - \frac{Fr}{R(r - r')}\right)\right\}\frac{(r - r')}{r}B.$$

The amount of gain per unit time by using LIFO is then

$$\frac{F}{R}\left(\frac{r'}{r - r'}A - B\right).$$

The situation in regard to Case I can be depicted as in Figure 3. The curve labeled a_1 corresponds to the same θ for both systems. The curve labeled a_2 corresponds to the same S for both systems.

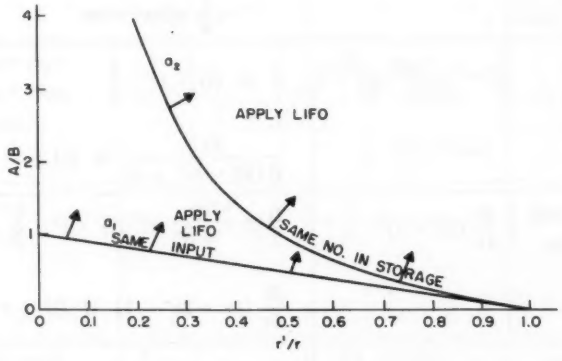


Figure 3 - Curves bounding LIFO areas

Any point above the appropriate curve has a pair of coordinates, that is, values of the two ratios, A/B and r'/r, such that on the latter comparison alone it is profitable financially to use LIFO. This does not include the possible monetary gains to be had under LIFO by a decreased incidence of breakdowns, crashes, associated malfunctions, etc., as well as intangibles like having the most up to date equipment in use, possible saving of life, and more certain accomplishment of mission. It is evident from Figure 3 that LIFO may be justified when only lower bounds on A/B and r'/r are known. Thus it appears that a minimum amount of information is required in order for one to decide whether to go to a LIFO system, namely knowledge that the deterioration curve of Figure 1 is not appreciably above the straight line, a lower bound of A/B and a lower bound of r'/r. The field of possible application of LIFO when the deterioration curve is concave down (the broken curve of Figure 1 is not closed. It is

perfectly possible that in some situations the gains in using LIFO would overbalance the cost of the additional input required to maintain the same number S in storage, but it has not been studied in this paper.

It is to be noted that the seven comparative criteria tabulated previously did not enter directly into the dollar comparisons which form the basis of the curves in Figure 3.

Case II: Assigned storage life is $r' \leq r$. Input of new items to storage is periodic, a constant number θ of them being introduced at the beginning of each unit time period.

Again the same set of comparative results under the two systems will be given as was done for Case I. See Appendices 2 and 3 for the derivations. The change from θ being given to S being given is mainly accomplished by solving (v) for θ as a function of S and substituting into the other elements. An interesting fact is that under FIFO when none go over age in storage (this is identified by (i) being less than $r' - \frac{1}{2}$), all the formulas of Case I are still true for Case II.

F, θ, r, r', R, A, B given. Periodic input.

	LIFO	FIFO
(i) average storage age of items at time they are sent to field	$\frac{1}{2}$	$r\left(1 - \frac{F}{\theta R}\right)$ if $< r' - \frac{1}{2}$ $r' - \frac{1}{2}$ otherwise
(ii) number of items over age in field per unit time	$\beta = \frac{2rF}{R(2r - 1)}$	θ if (i) $< r' - \frac{1}{2}$ $\frac{2rF}{R(2r - 2r' + 1)}$ if (i) $\geq r' - \frac{1}{2}$
(iii) average age of field items (in field life units)	$\frac{R}{4r}(2r + 1)$	$\frac{R}{2}\left(2 - \frac{F}{\theta R}\right)$ if (i) $< r' - \frac{1}{2}$ $\frac{R}{4r}(2r + 2r' - 1)$ if (i) $\geq r' - \frac{1}{2}$
(iv) input of new items per unit time	θ	θ
(v) average number of items maintained in storage	$S = \frac{\beta}{2} + r'(\theta - \beta)$	$S = \frac{r}{R}(\theta R - F)$ if (i) $< r' - \frac{1}{2}$ $S = r'\theta - \frac{rF}{R(2r - 2r' + 1)}$ if (i) $\geq r' - \frac{1}{2}$
(vi) rate of over age in storage per unit time	$a = \theta - \beta =$ $\theta - \frac{2rF}{R(2r - 1)}$	0 if (i) $< r' - \frac{1}{2}$ $\theta - \frac{2rF}{R(2r - 2r' + 1)}$ if (i) $\geq r' - \frac{1}{2}$

	LIFO	FIFO
(vii) average age of stored items	*	$\frac{r}{2}\left(1 - \frac{F}{\theta R}\right)$ if (i) $< r' - \frac{1}{2}$ $\frac{1}{4\theta}\left\{2\theta r' - \frac{2rF}{R(2r - 2r' + 1)}\right\}$ if (i) $\geq r' - \frac{1}{2}$

* $\frac{\theta + 2a + 3a(r'^2 - 1)}{3(\theta + a) + 6a(1 - \epsilon)[r' - 1] + 3a\epsilon^2[r']}$ which for $\epsilon = 0$ reduces to $\frac{\theta + 2a + 3a(r'^2 - 1)}{3\beta + 6a[r']}$.

If (i) $< r' - \frac{1}{2}$ under FIFO, LIFO is indicated if $\frac{A}{B} \geq \left(1 - \frac{r'}{r}\right)$. The amount of gain per unit time by using LIFO is $(\theta - \beta)\left\{A - \left(1 - \frac{r'}{r}\right)B\right\}$.

If (i) $\geq r' - \frac{1}{2}$ under FIFO, LIFO is indicated if $\frac{A}{B} \geq \left(1 - \frac{r'}{r}\right)$. The amount of gain per unit time by using LIFO is

$$\frac{4rF}{R}(r' - 1)\left\{A - \left(1 - \frac{r'}{r}\right)B\right\}.$$

F, S, r, r', R, A, B given. Periodic input.

	LIFO	FIFO
(i) average storage age of items at time they are sent to field	$\frac{1}{2}$	$\frac{rRS}{rF + RS}$ if $< r' - \frac{1}{2}$ $r' - \frac{1}{2}$ otherwise
(ii) number of items over age in field per unit time	$\beta = \frac{2rF}{R(2r - 1)}$	$\frac{S}{r} + \frac{F}{R}$ if (i) $< r' - \frac{1}{2}$ $\frac{2rF}{R(2r - 2r' + 1)}$ if (i) $\geq r' - \frac{1}{2}$
(iii) average age of field items (in field life units)	$\frac{R}{4r}(2r + 1)$	$\frac{R}{2}\left(\frac{Fr + 2SR}{Fr + SR}\right)$ if (i) $< r' - \frac{1}{2}$ $\frac{R}{4r}(2r + 2r' - 1)$ if (i) $\geq r' - \frac{1}{2}$
(iv) input of new items per unit time	$\theta = \frac{S + \beta(r' - \frac{1}{2})}{r'}$	$\frac{S}{r} + \frac{F}{R}$ if (i) $< r' - \frac{1}{2}$ $\frac{1}{r'}\left\{S + \frac{rF}{R(2r - 2r' + 1)}\right\}$ if (i) $\geq r' - \frac{1}{2}$
(v) average number of items maintained in storage	S	S

	LIFO	FIFO
(vi) rate of over age in storage per unit time	$a = \theta - \beta = \frac{2S - \beta}{2r'}$	0 if $(i) < r' - \frac{1}{2}$ $\frac{S}{r'} - \frac{(2r' - 1)}{r'} \frac{rF}{R(2r - 2r' + 1)}$ if $(i) \geq r' - \frac{1}{2}$
(vii) average age of stored items	Use (vii) of preceding tabulation.	$\frac{1}{2} \frac{rRS}{(rF + RS)}$ if $(i) < r' - \frac{1}{2}$ $\frac{1}{4\theta} \left\{ 2\theta r' - \frac{2rF}{R(2r - 2r' + 1)} \right\}$ if $(i) \geq r' - \frac{1}{2}$

If $(i) < r' - \frac{1}{2}$ under FIFO, LIFO is indicated if

$$\left\{ \frac{F}{R} + \frac{S}{r} - \frac{2rF}{R(2r - 1)} \right\} \frac{A}{B} \geq \frac{(r - r')}{rr'} \left\{ S - \frac{rF}{R(2r - 1)} \right\}.$$

If $(i) \geq r' - \frac{1}{2}$ under FIFO, LIFO is indicated if $\frac{A}{B} \geq \left(\frac{r}{r'} - 1\right)$, which is the simplification of

$$\left\{ \frac{2rF}{R(2r - 2r' + 1)} - \frac{2rF}{R(2r - 1)} \right\} A \geq \frac{(r - r')}{r} B \left\{ \frac{S}{r'} - \frac{\beta}{2r'} - \frac{S}{r'} + \frac{(2r' - 1)}{r'} \frac{rF}{R(2r - 2r' + 1)} \right\}.$$

The actual qualitative decision as to when it is financially profitable to use LIFO (ignoring the intangible considerations, and not using directly the seven tabulated characteristics) is generally rather simple. The considerations involved in making the decision will now be briefly summarized.

Case I. A new item is immediately available when needed.

F, θ, r, r', R, A, B are given and are the same for both systems.

$\frac{A}{B} \geq 1 - \frac{r'}{r}$ indicates LIFO.

Actually, knowing only that these seven qualities, F, θ, r, r', R, A, B are the same for both systems, and knowing lower bounds for the ratios $\frac{A}{B}, \frac{r'}{r}$ is sufficient to make the decision. One might then reject LIFO when he should have accepted it, but he will not accept it when he should have rejected it.

F, S, r, r', R, A, B are given and are the same for both systems.

$\frac{A}{B} \geq \frac{r}{r'} - 1$ indicates LIFO. Again the qualitative decision can be made knowing only that F, S, . . . , B are the same for both systems, and knowing lower bounds for the ratios $\frac{A}{B}, \frac{r}{r'}$.

Case II. Input of new items to storage is periodic.

F, θ, r, r', R, A, B are given and are the same for both systems.

$\frac{A}{B} \geq \left(1 - \frac{r'}{r}\right)$ indicates LIFO. It is evident that the same qualitative statement as above may be repeated here.

F, S, r, r', R, A, B are given and are the same for both systems.
This situation is more complicated than the preceding ones. If $\frac{rRS}{rF + RS} < r' - \frac{1}{2}$, then

$$\left\{\frac{F}{R} + \frac{S}{r} - \frac{2rF}{R(2r - 1)}\right\} \frac{A}{B} \geq \frac{(r - r')}{(rr')} \left\{S - \frac{rF}{R(2r - 1)}\right\}$$

indicates LIFO. Otherwise, if $\frac{rRS}{rF + RS} \geq r' - \frac{1}{2}$, $\frac{A}{B} \geq \left(\frac{r}{r'} - 1\right)$ indicates LIFO.

APPENDIX 1
DERIVATION OF FORMULAS FOR CASE I

LIFO

(i) Follows immediately from the requirement that a new item is available when needed, and the nature of the LIFO system.

(ii) Follows immediately from the definitions.

(iii) Just the average of the 0 and R field ages with which each issued item begins and ends.

(iv) Given.

(v) Assuming (vi) to be true, there must be as many items in each unit category of storage age as go over age per unit time. This equals $r'\left(\theta - \frac{F}{R}\right)$.

(vi) The input rate minus the rate issued to the field equals the rate over age in storage.

(vii) Under the assumptions, items in storage are extended uniformly over the possible ages 0 to r', with average $\frac{r'}{2}$.

FIFO

(i) Let X = average age of items at time sent to field. Then $\left(\frac{r - X}{r}\right)$ R = average remaining field life, and rate of over age in field = $\frac{F}{\frac{(r - X)R}{r}} = \frac{Fr}{R(r - X)}$. Now if the situation is such that none go over age in storage, it is evident that $X < r'$ and $\frac{Fr}{R(r - X)} = \theta$, the entire input. Solving, $X = r\left(1 - \frac{F}{\theta R}\right) < r'$. On the other hand, if some do go over age in storage, it is assumed that those issued are of storage age r'.

(ii) If (i) $< r'$, it is assumed that none go over age in storage, so the entire input θ goes over age in the field.
If (i) $= r'$, $\frac{(r - r')}{r}$ R is average field life remaining in each item at time it is sent to field. So rate of over age in field = $\frac{F}{\left(\frac{r - r'}{r}\right)R} = \frac{Fr}{R(r - r')}$.

(iii) (i) < r' implies $r\left(1 - \frac{F}{\theta R}\right)$ as average storage age of items at time they are sent to field. Translated into field age, i.e., from 0 to R, this is $R - \frac{F}{\theta}$ from which the average age of field items (in field life units) is $\frac{1}{2}\left(R + R - \frac{F}{\theta}\right) = \frac{R}{2}\left(2 - \frac{F}{\theta R}\right)$.

(i) = r' implies that this average field age is $\frac{1}{2}\left(R + \frac{r'}{r} R\right) = \frac{R}{2r}(r' + r)$.

(iv) Given.

(v) (i) < r' implies none go over age in storage. If the input is θ per unit time and the average age of items at time they are sent to the field is $\frac{r}{\theta R}(\theta R - F)$ or (i), then each unit time must have used up in storage $\theta\left\{\frac{r}{\theta R}(\theta R - F)\right\}$ storage life units. This can be accomplished only by having that many or $\frac{r}{R}(\theta R - F) = S$ items in storage.

(i) = r' implies that every item either goes over age on the shelf at age r' or is issued to the field at age r'. The same reasoning as used in (v) above shows that $S = \theta r'$ must be maintained in storage.

(vi) (i) < r' implies that none go over age in storage.

(i) = r'. Using (ii) above, the rate of over age on the shelf or in storage is evidently $\theta - \frac{Fr}{R(r - r')}$.

(vii) (i) < r'. This is just the average of 0 and (i).

(i) = r'. Obvicusly the average of 0 and r'.

APPENDIX 2
FIFO UNDER PERIODIC INPUT

Assume that none go over age in storage. It is clear that the condition for this to occur and also to be identified is that the average age of items at time they are sent to the field (i) is less than $r' - \frac{1}{2}$. Otherwise, all issues to the field from the oldest complete input are during the time interval r' - 1 to r' and any left over age in storage from that input would raise the average age beyond the midpoint of the time interval.

The assumption implies that all θ of any particular input are issued to field during some time t_1 to $t_1 + 1$. Thus of this input, θ, the average number in storage during t_1 to $t_1 + 1$ is $\theta/2$. From t_1 to $[t_1 + 1]$ there are $[t_1]$ whole sets of θ in storage preceding the set of age t_1. At time $[t_1 + 1]$, a new set of θ items is added to storage and there are then $[t_1 + 1]$ full "θ" sets to the end of $t_1 + 1$.

A weighted average gives the average number of full sets of θ during t_1 to $t_1 + 1$ to be

$$\left\{[t_1 + 1] - t_1\right\}[t_1] + \left\{1 + t_1 - [t_1 + 1]\right\}[t_1 + 1] = t_1$$

using $[t_1 + 1] = 1 + [t_1]$.

Thus the average number of items maintained in storage is

$$S = t_1\,\theta + \theta/2 = \theta\,(t_1 + \tfrac{1}{2})\,. \tag{1}$$

The average age of items at time sent to field is $t_1 + \frac{1}{2} = \frac{2t_1 + 1}{2}$.

The average remaining field life of those sent to field is $T = R - \frac{R}{r}\left(\frac{2t_1 + 1}{2}\right)$ from which the number required in the field per unit time equals

$$\frac{F}{R - \frac{R}{r}\frac{(2t_1 + 1)}{2}} = \frac{2rF}{R(2r - 2t_1 - 1)} = \theta\,. \tag{2}$$

(1) and (2) determine t_1 since either θ or S is given. Then $t_1 = \frac{2rRS - RS - rF}{2rF + 2RS}$ and $\theta = \frac{F}{R} + \frac{S}{r}$. The remaining formulas under this condition of FIFO follow readily from basic combinations of these results.

Now let (i) $\geq r' - \frac{1}{2}$, so that $u>0$ items go over age in storage per unit time. The argument analogous to the above goes as follows. All items issued to the field are between storage ages $r' - 1$ and r'. Due to the fact that new input to storage is made only at the beginning of each unit time period, the u items all go over age at the same time.

It is proposed to find first the average number S maintained in storage. It is convenient to count time momentarily from the instant a particular input of θ items is made. During time $r' - 1$ to $[r']$ there are $[r' - 1]$ whole sets of θ. At time $[r']$ a new set of θ is added to storage and there are then $[r']$ whole sets θ in storage from time $[r']$ to r'. A weighted average gives the average number of full sets θ to be $\{[r'] - (r' - 1)\}\,[r' - 1] + \{r' - [r']\}\,[r'] = r' - 1$, using $[r' - 1] = [r'] - 1$. Since, during each unit time period, one input of θ is undergoing depletion from θ to u, its average contribution to storage is $\frac{1}{2}(\theta + u)$. Therefore average S, or just

$$S = (r' - 1)\,\theta + \tfrac{1}{2}(\theta + u) = \tfrac{1}{2}(2r'\theta - \theta + u)\,. \tag{1}$$

Average age of items at time they are sent to field $= \frac{1}{2}(r' + r' - 1) = \frac{1}{2}(2r' - 1)$. Therefore, average remaining field life of these at that time is $T = R - \frac{R}{r}\left(\frac{2r' - 1}{2}\right)$. Then the number required to be sent to the field per unit time is

$$\frac{F}{R - \frac{R}{r}\left(\frac{2r' - 1}{2}\right)} = \frac{2rF}{R(2r - 2r' + 1)} = \theta - u\,. \tag{2}$$

Depending on whether S or θ is given, equations (1) and (2) take the useful forms

$$\begin{cases} \theta = \dfrac{1}{r'} \left\{ S + \dfrac{rF}{R(2r - 2r' + 1)} \right\} \\ u = \dfrac{1}{r'} \left\{ S - \dfrac{rF(2r' - 1)}{R(2r - 2r' + 1)} \right\}. \end{cases} \tag{3}$$

$$\begin{cases} S = r'\,\theta - \dfrac{rF}{R(2r - 2r' + 1)} \\ u \doteq \theta - \dfrac{2rF}{R(2r - 2r' + 1)} \ . \end{cases} \tag{4}$$

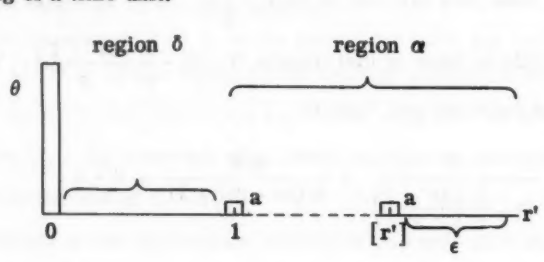
The only other item of the seven characteristics that might cause any difficulty in derivation is (vii). θ - u items are sent to the field per unit time. Their average storage age being $\frac{1}{2}(2r' - 1)$ at the time they are sent to the field, their average storage age is $\frac{1}{4}(2r' - 1)$. u items go over age in storage per unit time. Their average age in storage is $\frac{r'}{2}$. Thus the average storage age of all items is

$$\frac{(\theta - u)\frac{1}{4}(2r' - 1) + u\frac{r'}{2}}{(\theta - u) + u} = \frac{1}{4\theta}(2\theta r' - \theta + u) .$$

APPENDIX 3
LIFO UNDER PERIODIC INPUT

F, θ, r, r', R, A, B are given.

First there will be derived the average age of stored items, (vii), in the listing of elements or characteristics of the system. It will aid to note the following schematic diagram of storage at beginning of a time unit.



Let $0 \le t \le \epsilon$.

At time t,

number in $\alpha = [r']\,a$, average age $= \frac{1}{2}[r' + 1] + t$.

number in $\delta = (a - \theta)\,t + \theta$, age $= t$.

Therefore, average age at t is

$$\frac{a\,[r']\,(t + \frac{1}{2}[r' + 1]) + \{\theta + (a - \theta)\,t\}\,t}{a\,[r'] + \theta + (a - \theta)\,t}.$$

Now let $\epsilon \le t \le 1$.

At time t,

number in $\alpha = [r' - 1]\,a$, average age $= \frac{1}{2}[r'] + t$.

number in $\delta = (a - \theta)\,t + \theta$, age $= t$.

Therefore, average age at t is

$$\frac{a\,[r' - 1]\,(t + \frac{1}{2}[r'])\,+\,\{\theta + (a - \theta)\,t\}\,t}{a\,[r' - 1] + \theta + (a - \theta)\,t}.$$

As a general principle, if

N(t) = number of items in storage at time t,

A(t) = average age of items in storage at time t, then over $0 \le t \le 1$, the average age in storage is

$$\frac{\int_0^1 N(t)\;A(t)\;dt}{\int_0^1 N(t)\;dt}.$$

If the above functions are substituted into the latter expression (where one must use $\int_0^1 = \int_0^\epsilon + \int_\epsilon^1$ since the integrands are different in the two right hand integrals), it will reduce to

$$\frac{\theta + 2a + 3a(r'^2 - 1)}{3(\theta + a) + 6a(1 - \epsilon)[r' - 1] + 3a\epsilon^2[r']} = \text{average age of stored items.}$$

The above expression requires knowing a. Since $a = \theta - \beta$, β must be calculated. (i) is clearly $\frac{1}{2}$. Then average remaining field life is $\frac{R}{2r}(2r - 1)$ from which rate of overage in field per unit time is

$$\beta = \frac{F}{\frac{R}{2r}(2r - 1)} = \frac{2rF}{R(2r - 1)}.$$

(iii) is obtained as the average of R and $\frac{R}{2r}$.

(v) is obtained as follows:

During time 0 to ϵ, the average number in storage is $[r'] a + \theta + \frac{\epsilon}{2}(a - \theta)$, using some of the prior facts brought out in this appendix, and noting that the original θ goes down linearly to a in one time unit. During time ϵ to 1, the average number in storage is $[r' - 1] a + \frac{1}{2}\{a + \theta + \epsilon (a - \theta)\}$. A weighted average gives the average number in storage over the entire interval 0, 1 to be

$$S = \epsilon \{a [r'] + \theta + \frac{\epsilon}{2}(a - 1)\} + (1 - \epsilon) \left(a [r' - 1] + \frac{1}{2}\{a + \theta + \epsilon (a - \theta)\}\right).$$

By using $a = \theta - \beta$, $[r' - 1] = [r'] - 1$, $r' = [r'] + \epsilon$, this reduces to

$$S = \frac{\beta}{2} + r' (\theta - \beta).$$

* * *

CYCLING IN THE DUAL SIMPLEX ALGORITHM

E. M. L. Beale
Admiralty Research Laboratory,
Teddington, England

This note describes the construction of an example in which the usual iterative method for linear programming may fail unless special techniques for overcoming degeneracy are used.

A. J. Hoffman [3] has published an example of the fact that, when degeneracy occurs in the simplex algorithm for linear programming, the iteration may go into a closed loop if the variable to be made nonbasic is chosen arbitrarily from those that vanish at the next basic feasible solution. Despite the fact that no practical problem has yet been encountered where this "cycling" occurred, and that special devices are now available for preventing its occurrence, linear programmers are still intrigued by cycling and seek an understanding of the basic reasons underlying its occurrence. Unfortunately, the geometric motivation behind Hoffman's example is not easy to grasp. In this note it is pointed out that the geometrical interpretation of the changes of basis when degeneracy occurs is easier to follow in the dual method (due to C. E. Lemke [2]) than in the Simplex Method. This fact enabled me to construct an example of cycling in the dual algorithm, and the dual to this is an example of cycling in the simplex algorithm.

The relevant geometrical interpretation is present in Figure 1 of [2], provided that this figure is regarded as illustrating the dual to the problem being solved, since the simplex solution of the original problem there follows our interpretation of the dual method. The same geometrical interpretation is used in [1] in connection with the very similar method of leading variables. We now illustrate it with a simple numerical example. The numerical solution is written out in longhand, i.e., without using tableaux, in an attempt to make the process easier to follow. At each stage the basic variables, i.e., those associated with the vectors of the basis, are written on the left hand sides of the equations; the other, nonbasic, variables, which all take the value zero at the trial solution under consideration, are written on the right hand sides of the equations.

Suppose, then, that y_1, y_2, y_3, y_4 must be nonnegative and satisfy the equations

$$y_1 = -3 + y_3 + 2y_4,$$

$$y_2 = 5 - 2y_3 - 3y_4,$$

and that we wish to minimize

$$C = y_3 + 3y_4.$$

The dual solution is as follows:

$$y_3 = 3 + y_1 - 2y_4,$$

$$y_2 = -1 - 2y_1 + y_4,$$

$$C = 3 + y_1 + y_4;$$

so

$$y_3 = 1 - 3y_1 - 2y_2,$$

$$y_4 = 1 + 2y_1 + y_2,$$

$$C = 4 + 3y_1 + y_2.$$

Hence we should put $y_1 = y_2 = 0$, when $y_3 = y_4 = 1$ and $C = 4$.

This is analogous to the simplex solution to the dual problem, which can be written

$$x_3 = 1 - x_1 + 2x_2,$$

$$x_4 = 3 - 2x_1 + 3x_2,$$

minimize

$$C_x = -3x_1 + 5x_2.$$

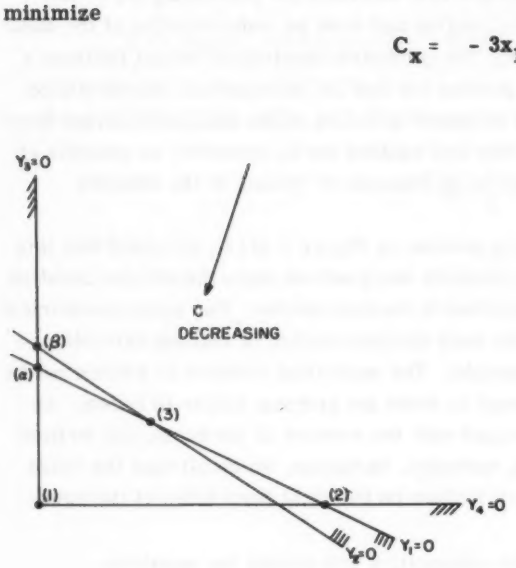


Figure 1

However, we wish to consider the geometrical interpretation of the original solution to the problem in the y_i. Figure 1 illustrates the lines $y_i = 0$ for $i = 1$ through 4, plotted with rectangular Cartesian co-ordinates y_3 and y_4. Each line is shaded on the side corresponding to negative, or inadmissible, values of the variable. A normal is also drawn to the lines C = constant, with an arrow pointing in the direction in which C decreases. By inspecting the figure we see that the only feasible solutions are represented by the points inside and on the triangle (3) (α) (β), and that the point (3) gives the smallest value of C. However our dual simplex algorithm starts at the point (1), where the original nonbasic variables y_3 and y_4 vanish. The algorithm requires that the coefficients of the nonbasic variables in the expression for C always remain nonnegative. Geometrically this means that the point where the nonbasic variables all vanish must give at least as small a value of C as any other point in the angle (in n dimensions the polyhedral cone) where the nonbasic variables are all nonnegative.

Now at the first trial solution y_1 is negative, i.e., the point (1) is on the wrong side of the line $y_1 = 0$. We remedy this by making y_1 nonbasic in place of one of the other nonbasic variables. This involves moving to either (2) or (γ). However we cannot move to (α) because this point does not give the smallest value of C inside its angle. (Algebraically the coefficient of y_3 is negative if C is expressed as a function of y_1 and y_3). The situation can most easily be expressed geometrically by saying that we must move to the point where C takes its smallest possible value consistent with the condition that all the old nonbasic variables, and the new one, must be nonnegative. (This implies that the new value of C cannot be less than the old one).

In this example our second point is (2). We then introduce the condition that y_2 must be non-negative as well as y_1 and y_4; this leads to the point (3), which represents the final solution.

Now degeneracy in the dual simplex method involves the vanishing of one or more of the coefficients in the expression for C in terms of the nonbasic variables. The value of C may then remain unchanged when the basis is changed, and, more fundamentally, there may be ambiguity in the choice of variable to be made basic. It is pointed out in [3] that this very rarely causes trouble in practice. For example in the problem represented in Figure 1 but with C made independent of y_1, y_2, y_3 and y_4, one can start the iteration from (1) and proceed to either (2) or (α). From (2) one must make y_3, the only negative variable, nonbasic, and this leads to (3) which is an acceptable final solution. If the first step led to (α) this is already an acceptable final solution.

We propose to show that by adding further lines (i.e., further variables) to the problem represented by Figure 1 with $C \equiv 0$, a cycle can be set up. However these additions must be made purposefully, so we first discuss the problem generally.

Now two successive steps in the iteration can never correspond to the same point in our geometrical representation, even when degeneracy occurs. One or more basic variables may vanish at a trial solution but only a variable that is actually negative can be made nonbasic at the next step. The cycle must therefore be represented by an actual closed loop in the diagram. Further, each move must be in a direction within the angle where the nonbasic variables are nonnegative. The obvious way to complete the loop is then to rotate this angle through 360°. A good start to this process has been made in Figure 1, and the complete loop is illustrated in Figure 2, where the variables y_3 and y_4 of Figure 1 have been renamed y_5 and y_6, two new variables y_3 and y_4 being introduced. The cycle is through the points (1), (2), (3), (4), (5), (6).

To construct a numerical example from the diagram, we must first choose co-ordinates for the points (1) through (6). Taking y_5 and y_6 as rectangular Cartesian co-ordinates, we can, for instance, put

$$(1) = (0, 0),\quad (2) = (3, 0),\quad (3) = (1, 1),$$
$$(4) = (-2, 3),\quad (5) = (-1, 1),\quad (6) = (0, -2).$$

This determines the expressions for y_1 through y_4 in terms of y_5 and y_6, up to 4 arbitrary positive factors, C_i. In fact

$$y_1 = C_1 (-3 + y_5 + 2y_6),$$
$$y_2 = C_2 (5 - 2y_5 - 3y_6),$$
$$y_3 = C_3 (-1 - 2y_5 - y_6),$$
$$y_4 = C_4 (2 + 3y_5 + y_6).$$

Now the normal procedure is to choose the variable that is most negative at any trial solution as the new nonbasic variable. We must therefore choose C_1 through C_4 so that this is always the variable we require for our loop.

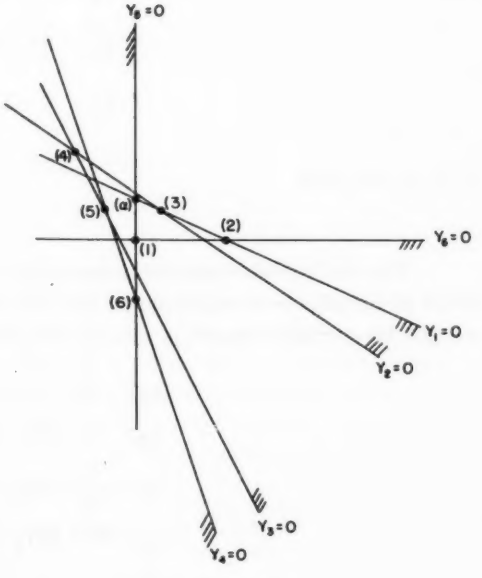


Figure 2

At the point (1), i.e., (0, 0), we require y_1 to be the most negative variable, so we must have

$$3C_1 > C_3.$$

Considering the other five points of the loop similarly, we obtain the four further conditions

$$C_2 > 7C_3,\ C_4 > 2,\ 1 > 2C_1,\ 2 > 7C_1.$$

All these inequalities are satisfied if, for example,

$$C_1 = \frac{1}{4},\ C_2 = 4,\ C_3 = \frac{1}{2},\ C_4 = 3.$$

From a theoretical point of view this is a valid example, but it has one unsatisfactory feature: the linear programming problem has no solution. This can be seen geometrically from the fact that in Figure 2 no point is on the right side of all the lines (or algebraically from the expression for y_3). However, it is clear that if the line $y_3 = 0$ is moved up so as to pass through the point (α) this will be a feasible solution to the problem. We therefore introduce a new variable y_7 proportional to the distance that the line $y_3 = 0$ is moved from its original position, and associate a positive coefficient in C with this variable. The problem is then to find nonnegative y_1 through y_7 satisfying

$$y_1 = -\frac{3}{4} + \frac{1}{4}y_5 + \frac{1}{2}y_6,$$

$$y_2 = 20 - 8y_5 - 12y_6,$$

$$y_3 = -\frac{1}{2} - y_5 - \frac{1}{2}y_6 + y_7,$$

$$y_4 = 6 + 9y_5 + 3y_6,$$

so as to minimize

$$C = y_7.$$

The dual simplex algorithm then proceeds as follows, assuming that when there is a choice of variable to be made basic, the one coming first in the tableau is chosen. (In this example the variable chosen is also the one with the lowest suffix).

$$y_5 = 3 + 4y_1 - 2y_6,$$

$$y_2 = 4 - 32y_1 + 4y_6,$$

$$y_3 = -\frac{7}{2} - 4y_1 + \frac{3}{2}y_6 + y_7,$$

$$y_4 = 33 + 36y_1 - 15y_6,$$

$$C = y_7,$$

so

$$y_5 = 1 - 12y_1 - \frac{1}{2}y_2,$$

$$y_6 = 1 + 8y_1 + \frac{1}{4}y_2,$$

$$y_3 = -2 + 8y_1 + \frac{3}{8}y_2 + y_7,$$

$$y_4 = 18 - 84y_1 - \frac{15}{4}y_2,$$

$$C = y_7;$$

so

$$y_5 = -2 - \frac{3}{2}y_3 + \frac{1}{16}y_2 + \frac{3}{2}y_7,$$

$$y_6 = 3 + y_3 - \frac{1}{8}y_2 - y_7,$$

$$y_1 = \frac{1}{4} + \frac{1}{8}y_3 - \frac{3}{64}y_2 - \frac{1}{8}y_7,$$

$$y_4 = -3 - \frac{21}{2}y_3 + \frac{3}{16}y_2 + \frac{21}{2}y_7,$$

$$C = y_7;$$

so

$$y_5 = -1 + 2y_3 + \frac{1}{3}y_4 - 2y_7,$$

$$y_6 = 1 - 6y_3 - \frac{2}{3}y_4 + 6y_7,$$

$$y_1 = -\frac{1}{2} - \frac{5}{2}y_3 - \frac{1}{4}y_4 + \frac{5}{2}y_7,$$

$$y_2 = 16 + 56y_3 + \frac{16}{3}y_4 - 56y_7,$$

$$C = y_7;$$

so

$$y_3 = \frac{1}{2} + \frac{1}{2}y_5 - \frac{1}{6}y_4 + y_7,$$

$$y_6 = -2 - 3y_5 + \frac{1}{3}y_4,$$

$$y_1 = -\frac{7}{4} - \frac{5}{4}y_5 + \frac{1}{6}y_4,$$

$$y_2 = 44 + 28y_5 - 4y_4,$$

$$C = y_7;$$

so

$$y_3 = -\frac{1}{2} - y_5 - \frac{1}{2}y_6 + y_7,$$

$$y_4 = 6 + 9y_5 + 3y_6,$$

$$y_1 = -\frac{3}{4} + \frac{1}{4}y_5 + \frac{1}{2}y_6,$$

$$y_2 = 20 - 8y_5 - 12y_6,$$

$$C = y_7.$$

The cycle is now complete.

The same cycle arises if the simplex method is applied in the same way to the dual problem, which can be written

$$x_5 = -\frac{1}{4}x_1 + 8x_2 + x_3 - 9x_4,$$

$$x_6 = -\frac{1}{2}x_1 + 12x_2 + \frac{1}{2}x_3 - 3x_4,$$

$$x_7 = 1 - x_3,$$

minimize

$$C_x = -\frac{3}{4}x_1 + 20x_2 - \frac{1}{2}x_3 + 6x_4.$$

Of course either problem can be solved by using one of the ϵ-perturbation schemes that have been devised for avoiding degeneracy - or indeed by simply choosing the new variable at random, since the solution follows inevitably if the alternative choice of new variable is made at any of the three points in the cycle where a choice exists. Incidentally the solution is

$$C = \frac{5}{4},\ y_2 = 2,\ y_4 = \frac{21}{2},\ y_6 = \frac{3}{2},\ y_7 = \frac{5}{4},\ y_1 = y_3 = y_5 = 0,$$

or

$$C_x = \frac{5}{4},\ x_1 = 1,\ x_3 = 1,\ x_5 = \frac{3}{4},\ x_2 = x_4 = x_6 = x_7 = 0.$$

Finally we consider briefly whether any general conclusions can be drawn from this work about the circumstances in which such cycles can occur.

In the dual method no trial solution with only one nonbasic variable having a zero coefficient in the expression for C can be part of a cycle: for the problem is then effectively in only one dimension, and the trial solution can move in only one direction. The same applies to the simplex method with only one basic variable vanishing at the feasible solution.

Further, in the dual method no trial solution with just two nonbasic variables having a zero coefficient in the expression for C can be part of a cycle containing less than six steps. For we can imagine a point travelling round the cycle. The path of this point consists of straight line segments, the direction of motion changing only at a trial solution. Further, since the possible new directions in which the point can move lie in the angle between the present line segment of the path and the previous one produced, the change in direction must be first to

the left, then to the right, then to the left and so on. It follows that the number of turning points must be even, and, as is clear from Figure 3, that the first and fifth such points cannot coincide, since they must be on opposite sides of any line crossing the path at the third point. Therefore the number of steps in the cycle must be even and not less than five, in fact at least six.

An analysis of the situation in three or more dimensions appears to require a more sophisticated approach, and will not be attempted here.

I am grateful to the Admiralty for permission to publish this paper.

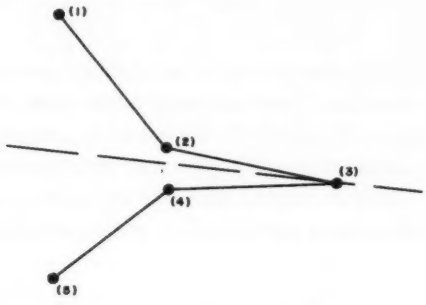


Figure 3

REFERENCES

[1] E. M. L. Beale. "An Alternative Method for Linear Programming," Proc. Camb. Phil. Soc. 50 pp. 513-523 (1954).

[2] C. E. Lemke, "The Dual Method of Solving the Linear Programming Problem," Naval Research Logistics Quarterly 1 pp. 36-47 (1954).

[3] A. J. Hoffman, "Cycling in the Simplex Algorithm," unpublished N.B.S. Report.

* * *

COMPUTATION OF MAXIMAL FLOWS IN NETWORKS

D. R. Fulkerson and G. B. Dantzig
RAND Corporation

A simple computational method, based on the simplex algorithm of linear programming, is proposed for the following problem:

"Consider a network (e.g., rail, road, communication network) connecting two given points by way of a number of intermediate points, where each link of the network has a number assigned to it representing its capacity. Assuming a steady state condition, find a maximal flow from one given point to the other."

The linear programming formulation of the network-flow problem given in [1] is not only useful as a theoretical tool: when suitably interpreted, it provides in addition a simple and efficient hand-computing scheme. It is our aim in Section I to describe this computation for networks with capacity constraints on arcs only. It is unnecessary to use programming terminology in this description, and so we dispense with it. A direct proof will be given in Section II that the computational method assures one of finding a maximal flow in an arbitrary network.

DESCRIPTION OF METHOD

The problem (see [1], [2]) may be stated as follows: one is given a connected network of arcs and nodes with two distinguished nodes, called source and sink respectively. All other nodes are called intermediate. Each arc in the network has associated with it a positive integer, its flow capacity. The direction of flow is assumed to be away from the source in arcs having the source as endpoint, and into the sink in sink arcs; otherwise no flow direction is specified. Subject to the conditions that the flow in an arc does not exceed its capacity, and that the total flow into any intermediate node is equal to the flow out of it, it is desired to find a maximal flow in the network, i.e., a flow which maximizes the sum of the flows in source (or sink) arcs.

In order to illustrate the computing method, let us consider a simple example. Suppose we have the network of Figure 1 with source A, sink B, and arc capacities as indicated. To start out, select two trees[1] of arcs, one branching out from A, the other from B, so that every intermediate node is reached by just one of the trees. Call these T_A and T_B. For example, T_B might consist of B alone, and T_A might be the set AC, CD, DE, EF (see Figure 1a). Notice that since the network is connected, it is always possible to select two such trees.[2] Next introduce any arc which leads from T_A to T_B. There will then be just one chain from A to B; flow

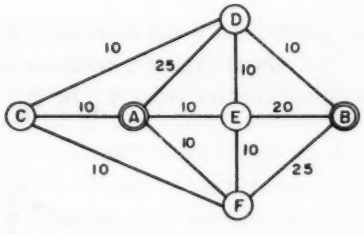


Figure 1

[1] A tree is a connected linear graph without circuits; that is, there is one and only one chain of arcs joining each pair of nodes.

[2] It is clear that arcs may be removed until a tree is left. There is then a unique chain joining A and B. Elimination of any arc of this chain gives two trees of the kind described.

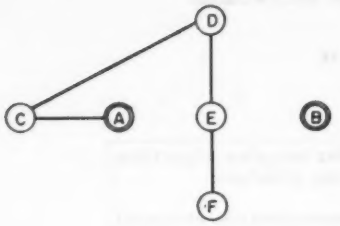


Figure 1a

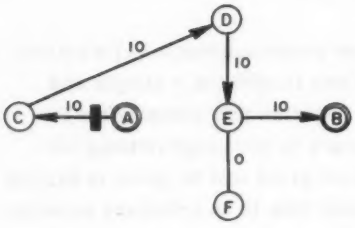


Figure 2

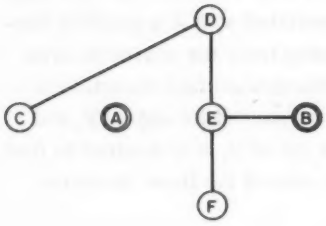


Figure 2a

as much as possible along this chain. In the example, EB is such an arc, and we have then the flow diagram of Figure 2 with all of the arcs AC, CD, DE saturated. Select any one of these saturated arcs, say AC, and place some identifying mark on it for future reference. In Figure 2 we have used a bar (▮); this symbol will be used throughout to designate a subset of the saturated arcs. Now observe that if the barred arc is dropped from the picture, we again have two trees (Figure 2a) $T_A = \{A\}$ and $T_B = \{EB, ED, DC, EF\}$. Again introduce any unbarred arc leading from T_A to T_B, say AF. This creates a flow along the chain AF, FE, EB of 10 units and saturates each arc of this chain. Select one of these, say AF, and bar it. We now have the following diagram (Figure 3), and we have achieved a flow of 20. Dropping barred arcs gives the same trees as in Figure 2a. Introduce arc AE from T_A to T_B. This leads to a situation we have not met previously in that the chain thus constructed, namely AE, EB, can not take any more flow because EB, though unbarred, is at its upper limit. Bar EB and leave AE in with a flow of zero, obtaining Figure 4 with new trees as shown in Figure 4a. Introduce DB to get the chain AE, ED, DB. This time we can get an increase even though DE is saturated, since the flow in Figure 4 is from D to E. Thus if the flow from A to E is increased by $\epsilon \geq 0$, the flow from D to E must be decreased by ϵ, and the flow from D to B increased by ϵ (see Figure 4b) in order to preserve the conservation equations at E and D. The largest possible value of ϵ is 10, since the capacity of DB (and of AE) is 10. This cancels the flow from D to E. Bar DB and proceed.

A repetition of steps of the kind described produces the sequence of flows depicted in Figures 5 through 8 below.

Now in Figure 8 the trees are $T_A = \{AD, DC\}$, $T_B = \{FB, EF\}$, and there are no more arcs to introduce from T_A to T_B. At this stage examine the barred arcs connecting nodes of

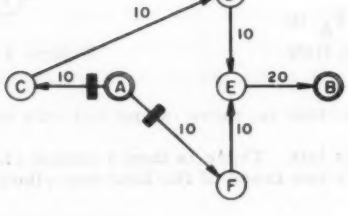


Figure 3

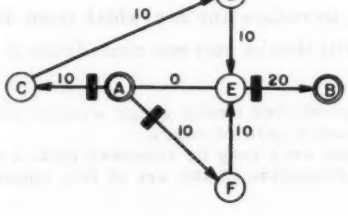


Figure 4

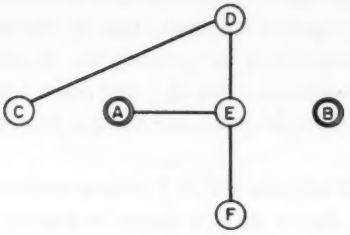


Figure 4a

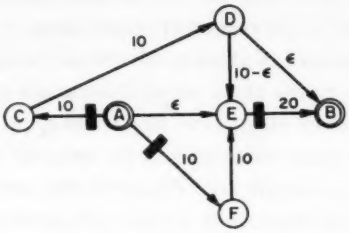


Figure 4b

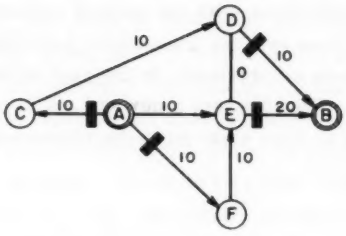


Figure 5

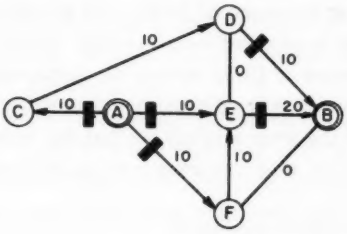


Figure 6

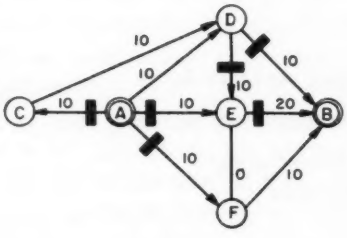


Figure 7

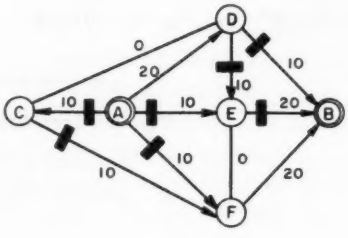
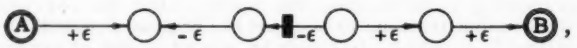


Figure 8

T_A to those of T_B. If the flow in each of these is in the "right" direction, that is, from T_A to T_B, an optimum has been reached, because these arcs now constitute, in the original network, a cut separating A and B (i.e., every chain joining A and B contains one of these arcs), and the total flow through the network is equal to the sum of the capacities of the arcs forming the cut. Hence, since it is clear that no flow can exceed the sum of capacities of arcs forming a cut, the flow is maximal in this case. If, on the other hand, the flow in one of the barred arcs which join T_A to T_B is in the wrong direction, an increase in total flow may possibly be obtained by decreasing the flow in this arc. To see this, notice that the arc in question, together with arcs of T_A and T_B, will form a (unique) chain joining A and B which might look, for example, like this



the arrowheads denoting directions of flow. Now we may subtract $\epsilon \geq 0$ from the flow in all arcs having the same direction as the barred arc, add ϵ to the flow in the other arcs.

Since these latter arcs will of necessity include the source (and sink) arc in the chain, the total flow will be increased provided there is any slack in the chain, that is, if none of the arcs where $+\epsilon$ appears is already saturated (though unbarred in our procedure). In any case, locate an arc in the chain which determines the maximum value of ϵ and place a bar on it. This either gives rise to two new trees or, if the position of the bar has not been changed, reverses the flow completely in the original arc.

In the example (see Figure 8) the arcs DB, DE, AE, AF, CF joining nodes of T_A to those of T_B all have flows in the right direction. Hence the flow shown in Figure 8 is maximal, and consequently these arcs form a minimal cut, i.e., a cut the sum of whose arc capacities is minimal.

It is clear from the preceding discussion that the proposed computation provides a nondecreasing sequence of flows. All practical experience with the method indicates that a maximal flow is reached very quickly. There is, however, the theoretical possibility that an infinite sequence of flows having the same value may be obtained. In the rest of the paper we shall indicate one way of removing this difficulty. The idea, not a new one, is to perturb the capacities of the arcs slightly so that the decision at each state as to the placement of the bar becomes unique, thus assuring a finite process.[3]

PROOF OF CONVERGENCE

Given an arbitrary network N, let $a_1, \ldots, a_n$ be a listing of the arcs in some order and let $c_1, \ldots, c_n$ be their capacities.

At each stage of the computation, the arcs a_i, $i=1, \ldots, n$, are divided into four mutually exclusive classes T_A, T_B, S, Z where

T_A is a tree branching out from A,

T_B is a tree branching out from B,

S consists of the barred arcs,

Z contains all other arcs,

and each node is an endpoint of some arc of T_A or T_B. Call such a division a basic partition of N.

Theorem

For a given basic partition of N, let arbitrary flows be assigned to arcs of S and Z. Then the conservation equations uniquely determine flows in the arcs of T_A and T_B. Moreover, if $x_1, \ldots, x_m$ are the amounts of the assigned flows, then the amount of flow in each arc of N is equal to $\sum_{j=1}^{m} \delta_j x_j$, where $\delta_j = 1, 0, \text{or } -1$.

The proof can be made clear by considering the following example where the dotted flows $x_1, \ldots, x_8$ are assigned.

---

[3] From the linear programming point of view the method thus far corresponds to the original simplex algorithm without modifications to account for degeneracy, with this difference. When "ties" occur in the method (i.e., when a bar may be placed in one of several positions at some step), we have not indicated a unique choice, but rather have used language like "pick any one..." If we interpret this to mean "choose at random among the possibilities," then we are at least assured that with probability one the computation terminates, as we shall see.

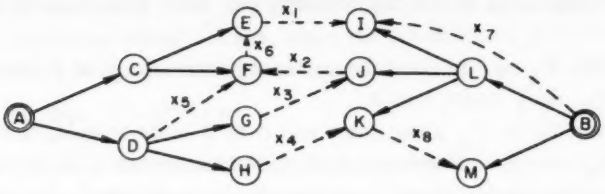


Figure 9

Orient each arc of the two trees in some fashion, say away from A in T_A and away from B in T_B (as shown in Figure 9), the convention being that a positive number on an arc represents flow in the same direction as its orientation. Thus if x, y, z are the flows in arcs AC, CE, CF, respectively, the equation at C is $x - y - z = 0$.

That the flows in the arcs of each tree are uniquely determined can be seen by solving the equations recursively in the following manner. Start with an outermost node in one of the trees, E for example. There is then only one unknown flow into E, hence its value can be found; here it is $x_1 - x_6$. Similarly determine the flows in all arcs branching out of F, G, H, and then work backward in the tree.

For the last assertion of the theorem, it is convenient to solve the equations in another way. Select any one of the assigned flows x_i. There are two cases to consider. Either the arc with flow x_i joins two nodes of the same tree or not. Suppose it joins two nodes of T_A, as EF does, for instance. Since there is a unique chain joining E and F in T_A, namely FC, CE, the flow x_6 in EF can be taken care of by assigning $\pm x_6$ appropriately to the arcs of this chain. Here the assignment is $+x_6$ to CF, $-x_6$ to CE. On the other hand, if the arc in question joins T_A to T_B, as x_1 does, consider the chain joining A and B and having this arc as one of its links. Here it is AC, CE, EI, IL, LB, and the assignments are x_1, x_1, x_1, $-x_1$, $-x_1$, respectively. If we continue in this fashion, it is evident that a solution to the system of equations will be built up having the form described in the theorem. For the example, the solution is shown in Figure 10.

In the algorithm outlined in the preceding section, arcs of S are saturated and arcs of Z have zero flow. Hence, we may state

Corollary

Suppose S consists of arcs $a_{i_1}, \ldots, a_{i_s}$ at any given step. Then the flow in each arc of N is given by a sum of the form $\sum_{k=1}^{s} \delta_{i_k} c_{i_k}$, where $\delta_{i_k} = 1, 0, \text{or } -1$.

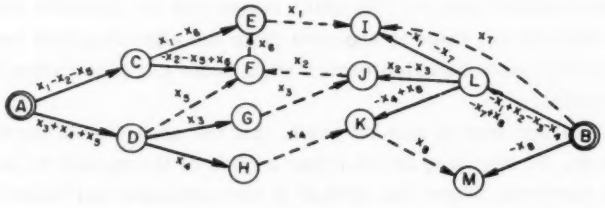


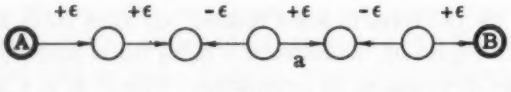
Figure 10

Let us examine in more detail how ties occur in the decision as to which arc to put into S in going from one step to the next in the computation. Each step is of one of the following two types:

(a) a flow from T_A to T_B is introduced along some arc $\underline{a}$ of Z, thus shifting one of the arcs of T_A or T_B, or $\underline{a}$ itself, into S;

(b) a flow from T_B to T_A along some arc b of S is decreased, again shifting one of the arcs of T_A or T_B into S, or leaving b in S with its direction of flow reversed.

It suffices to consider (a), since (b) is essentially no different. We have then a chain from A to B



consisting of $\underline{a}$ and arcs of T_A and T_B, with flows in one direction or the other in each arc of the chain, the direction of flow in the source (sink) arc being away from A (into B). Suppose the arcs in the chain are $a_{j_1}, a_{j_2}, \ldots, a_{j_m} = a, \ldots, a_{j_r}$ and let S consist of $a_{i_1}, \ldots, a_{i_s}$. By the corollary, the magnitude of flow in a_{j_ℓ} is given by $\sum_{k=1}^{s} \delta_{i_k}^{j_\ell} c_{i_k}$. It follows that we may increase the flow in arc $\underline{a}$ by an amount

$$\epsilon = \min_{j_\ell} \left(c_{j_\ell} \pm \sum_{k=1}^{s} \delta_{i_k}^{j_\ell} c_{i_k} \right)$$

where the plus sign is taken if a_{j_ℓ} has flow directed toward A, the minus sign otherwise. Since the c_{j_ℓ} are integers, it is clear that if, instead of using the original capacities, we had altered them to begin with by setting

$$c_1' = c_i + 10^{-i}, \quad i = 1, \ldots, n,$$

the minimum would be achieved for just one j_ℓ. Then at no stage do ties occur, and consequently arcs not in S will, at each step, be unsaturated. This means that the total flow through N is increased with each iteration, and hence, by the theorem, no basic partition having the same directions of flow in arcs of S can reoccur. Since there are only a finite number of basic partitions, the procedure terminates for the perturbed problem. Thus, we must arrive at a flow in which all arcs of S leading from T_A to T_B form a cut and are in the right direction; hence, the flow is maximal. Once an optimum has been attained, a solution to the original problem is given by rounding each arc flow to the nearest integer.

Notice that the theorem (see [1], [2]) which states that the maximal flow value in a network is equal to the minimal cut value is apparent from the algorithm just outlined, since the point at which the computation terminates occurs when some cut in the network is saturated by flows in the right direction.

We also point out, for what it may be worth, that the assertion in the footnote on page 280 can now be verified. To see this, let us define a state of the system to be any flow corresponding to a basic partition, where the arcs of S are saturated and those of Z have zero flow, i.e., a state is any flow that might conceivably be reached in the computation. There are

only a finite number of states, say $\sigma_1, \ldots, \sigma_t$. Thus, if we start with any state σ_i and randomize among ties at each step, there is a positive probability p_i of reaching a state satisfying the maximality criterion within t steps, since the possibilities at each step include the one which would be selected in the perturbed problem. Let $p = \min_i p_i > 0$. Then the probability of not terminating is $\lim_{m \to \infty} (1 - p)^m = 0$.

For anyone who may be interested in computing maximal flows in networks by this method, we do not recommend altering the capacities as described in this section. All empirical evidence, both in this particular problem and with the simplex algorithm in general (of which this is a special case), indicates that the insurance bought in this way is not worth the effort.

REFERENCES

[1] Dantzig, G. B., and D. R. Fulkerson, "On the Min Cut Max Flow Theorem of Networks," to appear in "Papers on Linear Inequalities and Related Systems, edited by H. W. Kuhn and A. W. Tucker, Annals of Mathematics Study 38.

[2] Ford, L. R., and D. R. Fulkerson, "Maximal Flow through a Network," to appear in Canadian Journal of Mathematics.

* * *

THE STRUCTURE OF DYNAMIC PROGRAMING MODELS

Samuel Karlin
California Institute of Technology

This paper discusses the structure and formulation of dynamic programing models. R. Bellman's "Principle of Optimality" is reexamined for deterministic models. A formulation of the general stochastic dynamic programing model is given and applied to an inventory model.

In recent years much attention has been devoted to the study of multistage decision problems or alternately known as dynamic programing models. The example of inventory analysis presented by Arrow, Harris, and Marschak exhibits one such important dynamic model [1]. Bellman has introduced several other instances in a series of publications [2], [3]. In view of these and other references cited later and the growing importance both practically and theoretically of such dynamic models, it would be of interest to delimit precisely the structure of dynamic programing models.

The principal difference between these multistage models and the static models treated so far is that the procedures (= policies = strategies) available in the static circumstance are much simpler to describe. A strategy is selected once and for all to be carried out, while the strategies available in the dynamic case are generally complicated functions of the information and actions undertaken in the preceding stages. Of course, by taking the proper global point of view and defining a strategy appropriately, any dynamic model can always be reduced to a static situation. This is a simple exercise in use of the tools of abstract mathematics. In general however, this so complicates the nature of the strategy that it is often impossible to analyze the problem in terms of this reduction. Alternately, the mode of procedure has been to express the outcome or yield of any strategy by a recursion formula. The possibility of employing recursion formulas requires a certain infinite symmetry to the model. This is fulfilled for instance if the total payoff is the sum of similar stagewise payoffs. This covers a variety of circumstances which includes special cases as where the stagewise yield corresponds to expected value or probability of accomplishing an objective on a given trial. The yield according to the recursion expression takes the form: the total yield according to a given policy is the return from the first stage plus the cumulative return of all later stages pursuing the specified policy. Arguing along the same lines, a special functional equation is then deduced for the optimal policy. This last equation is then carefully analyzed in an attempt to determine the optimal policies which represent the sought for solutions to the problem. The derivation of this functional equation referred to by Bellman as the "Principle of Optimality" [2] often involves some tacit assumptions and requires some care in operating with it. The underlying assumptions describing the model are often vague and the nature of what constitutes strategies is frequently not made clear.

It is the purpose of this paper to build a basis from which dynamic models can be examined. The spirit of this approach is related to Wald's formulation of statistical decision theory. The flexibility of the formulation presented below is large. We do not intend to imply

that our formulation of dynamic models encompasses all possible such problems which can arise. Nevertheless, it is felt that the spirit of this analysis is fundamental in the study of dynamic models and will help in one's study of particular examples.

In Section 1 the structure of the deterministic dynamic programing model is set up with a derivation of the "Principle of Optimality" coming out as a consequence. Some examples introduced by Bellman are then cast into this framework. The second section is concerned with a formulation of the general stochastic dynamic programing model. As an example, the inventory model is then treated from this point of view.

DETERMINISTIC DYNAMIC PROGRAMING MODEL

We begin our discussion by presenting a brief informal description of the deterministic dynamic programing process. A repeated number, usually infinite, of decisions must be made. As a result, a total cumulative yield is secured, which is the sum of the contributions of the individual steps in the model. The objective is to find that mode of procedure which maximizes the total value. The stagewise contributions according to any strategy is generally dependent upon the previous decisions and the random or predictive phenomenon affecting the process. A strategy, of course, is defined as a complete prescription how to act at each stage taking into account all contingencies. It follows that in the deterministic case that a strategy can be described as an infinite sequence of decisions to be made at each stage which is prescribed in advance. Having highlighted the principle ingredients and objectives of the deterministic process, we now direct attention to a detailed examination of the problem.

FORMAL STRUCTURE

In the deterministic model there are two essential spaces. Ω is the state space, the space of all possible situations in any given stage which can confront us. D is the decision space. It is the totality of all possible actions we could take in a given stage. A specific decision, an element of D, will be denoted by δ. Defined on the product space $\Omega \times D$ is a non-negative function $L_n(\omega, \delta)$ which for each n gives the return to us if at stage n we are faced with state ω and make the decision δ. Corresponding to each δ there is defined a transition operator T_δ which maps Ω into itself. After having made decision δ when faced with the state ω, then $T_\delta \omega = \omega'$ represents the new state which will confront us at the next stage. The mapping T_δ embodies the mechanism which transforms a given state into a new state after a decision has been made.

The following assumptions are imposed on these spaces:

(1) Ω is a topological Hausdorff space.

(2) D is a compact Hausdorff space.

(3) $L_n(\omega, \delta)$ is a real valued function, continuous on the product space $\Omega \times D$.

(4) For each δ, $T_\delta \omega$ is a continuous function of ω and for each ω, $T_\delta \omega$ is a continuous function of δ.

The strategy or policy space S is defined as the infinite product space of D replicated: $S = D \times D \times \ldots$. An element s in S is $s = (\delta_1, \delta_2, \delta_3, \ldots)$ where δ_1 is the decision made at the first stage, δ_2 the decision at the second, and so on. The reason that a strategy for the whole dynamic problem can be so described is a consequence of the deterministic nature of the model. For any given initial state ω and any decision δ_1 the nature of the next state $\omega' = T_\delta \omega$ is completely predictable in advance as a function of any decision involved previously.

There is defined a sequence of real positive functions $\{P_n(s)\}$ on S. The n^{th} function $P_n(s)$ is a function which depends only on the first n-1 components of s ($P_1(s) \equiv 1$) and shall be assumed to be continuous in s. These functions play the role of normalization constants for the ultimate yield. Usually $P_n(s)$ will be of the form $P_n(s) = \prod_{i=1}^{n-1} P(\delta_i)$ where $s = (\delta_1, \delta_2, \dots)$ and $P_1(s) \equiv 1$. In fact, unless stated to the contrary we take $P_n(s)$ of the product form as indicated.

The total yield is now defined for a given initial state ω and strategy s as

$$\Phi(\omega, s) = \sum_{n=1}^{\infty} L_n(\omega_n, \delta_n) P_n(s)$$

where $\omega_n = T_{\delta_{n-1}} \omega_{n-1}$ and $\omega_1 = \omega$. In most cases all the L_n's will be equal to a single function L, i.e., $L_n(\omega_n, \delta_n) = L(\omega_n, \delta_n)$ for all n.

The objective is to maximize the payoff function Φ for each initial state ω. Any strategy s for which $\Phi(\omega, s)$ achieves its maximum (if such an s exists) is called an optimal strategy. First, we determine sufficient conditions for the existence of optimal strategies. Indeed, by a classical theorem of Tychonoff on the direct product of compact spaces, the strategy space is compact since D is compact. Thus, if Φ were a continuous function on S for each ω it would follow that Φ achieves its maximum on S for each ω and thus optimal strategies exist.

To guarantee that this is the case we shall impose a fundamental requirement. First, note that each $L(\omega, \delta) P_n(s)$ is a continuous function of s. Our basic assumption is now that

$$\sum_{n=1}^{k} L(\omega_n, \delta_n) P_n(s) \tag{1}$$

converges uniformly in s. In view of this assumption of uniform convergence Φ is a continuous function of s. Hence, it achieves its maximum, as Φ is defined on the compact space S for each ω.

$$(\max_{S} \Phi(\omega, s) = \Phi(\omega, s^*)).$$

Summarizing,

Theorem 1. An optimal strategy s^* exists provided the series in (1) converges uniformly in S for each ω.

Examples

We shall now discuss briefly the following two examples in terms of our general formalism [2].

Example 1. Associated with two operations A and B are the following specified numbers:

A	B	Operations
x	y	Initial total worths
$0 < p_1 < 1$	$0 < p_2 < 1$	Probabilities of continuing
$0 < s_1 \leq 1$	$0 < s_2 \leq 1$	Proportions of return of successful operation

If operation A is performed p_1 is the probability that the operation is successful and another operation can be performed at the next stage. If the operation A is successful then the return or yield is the proportion s_1 of the initial total worth x. The maximum total worth that can be gained through repeating A at the next stage is reduced to $(1 - s_1)x$ so if A is repeated again successfully the return is $s_1(1 - s_1)x$. If, however, the operation is unsuccessful (with probability $1 - p_1$), there is no return from the operation nor can A or even B be repeated again. Similar interpretations hold for the quantities y, p_2, and s_2 of B. It is to be noted again that if at some stage an operation is unsuccessful then all operations stop.

The only nondeterministic feature to this model is contained in the probabilistic nature of the termination of the process. In spite of this probability aspect, it is easy to see in this case that strategies can be viewed as sequences of decisions. The reader can easily convince himself that this model is very much in the spirit of a deterministic model and consequently, Theorem 1 can be applied.

A physical interpretation of this model provided by R. Bellman is the following: there are two mines A and B with respective values x and y. Both mines cannot be operated simultaneously since there is only one mining machine. If the machine is used on mine A for a unit period of time the proportion s_1 of the total value x of A is mined. If the machine is used on B, the return in one unit of time is s_2y. Operations might not go on forever, however, since there is a probability $1 - p_1$ that the machine breaks down if it is used on A and a probability $1 - p_2$ that it breaks down if it is used on B. We search for the optimal procedure of selecting operations A and B so as to maximize the total expected yield.

A generalized form of this model involving several operations instead of only two shall now be described.

We identify Ω = a bounded closed portion of E^m (Euclidean space), with the usual topology imposed on Ω. D consists of a collection of m points which are labeled $(\delta_1, \delta_2, \ldots, \delta_m)$. That is, at each stage there are only m different possible decisions corresponding essentially to the choice of which operation is made. The topology on D is the discrete topology which is trivially compact. $L(\omega, \delta_i) = s_i \omega_i$ where $\omega = (\omega_1, \ldots, \omega_m)$ is a point in Ω and $s_1, s_2, \ldots, s_m$ is a set of constants. T_{δ_i} is the operator

$$(\omega_1, \ldots, \omega_m) \rightarrow (\omega_1, \ldots (1 - s_i)\omega_i, \ldots, \omega_m)$$

where the constants s_i are so restricted that T_{δ_i} maps Ω into Ω. The normalization function is taken as $P(\delta_i) = p_i$ $(0 < p_i < 1)$.

We now check whether the series in (1) converges uniformly. Since Ω is a bounded portion of E^m there exists an M such that $L(\omega, \delta) \leq M$ for all ω and δ. Since there are only a finite number of p_i's and they are all < 1, there exists an $\alpha < 1$ such that $p_i < \alpha \leq 1$ for all i. Hence,

$$\sum_{n=1}^{k} L(\omega_n, \delta_n) \prod_{i=1}^{n-1} P(\delta_i) \leq M \sum_{i=1}^{k} \alpha^{i-1}$$

as $\sum_{n=1}^{\infty} \alpha^{n-1}$ is convergent for $\alpha < 1$, $\sum_{n} L(\omega_n, \delta_n) \prod_{i=1}^{n} P(\delta_i)$ is uniformly convergent and therefore by Theorem 1 an optimal strategy exists.

In contrast, if some $p_i = 1$, then no optimal strategies exist although the yield corresponding to any strategy is finite. In fact, a uniform bound exists for the yield of any strategy.

Example 2. There is a total of x_0 resources and two operations A and B. If we divide the total resources into y which we invest in A and x_0-y which we invest in B, the total return is $g(y) + h(x_0-y)$, $g(y)$ coming from A and $h(x_0-y)$ from B. The amounts of resources remaining are ay from A and $b(x_0-y)$ from B so at the next stage our total resources are $ay + b(x_0-y)$, which we can redivide into operations A and B. This process continues until we have no resources left. The objective is to maximize the total return.

As the problem stands now we have no guarantee that the total return will have an upper bound. Let us assume that g and h are continuous monotonic increasing, $g(0) = h(0) = 0$, and $g'(0) < \infty$, $h'(0) < \infty$. Take Ω = positive real line with the usual topology. Set $D = \{\delta : 0 \leq \delta \leq 1\}$. Here again the topology is the usual one and D is compact. The yield for the state of nature ω and decision δ is given by $L(\omega, \delta) = h(\delta\,\omega) + g(1 - \delta)\,\omega)$. The transition operator is defined as $T_\delta\,\omega = b(\delta)\,\omega = a\,\delta\,\omega + b(1 - \delta)\,\omega$ where $a, b \leq c < 1$. Put $P_n(s) = 1$ for all s. It will be assumed also that there exists a $0 < c < 1$ such that a, b < c and

$$\sum_{n} g(c^n x_0) < \infty \text{ and } \sum_{n} h(c^n x_0) < \infty.$$

From this we conclude that any procedure yields a finite return. The total yield becomes

$$\Phi(\omega, s) = h(\delta_1\,\omega) + g(1 - \delta_1)\,\omega) + h(\delta_2\,b(\delta_1)\,\omega) + g((1 - \delta_2\,b(\delta_1)\,\omega)$$

$$+ h(\delta_3\,b(\delta_2)\,b(\delta_1)\,\omega) + \ldots\ldots\ldots$$

$$= \sum_{n=1}^{\infty} L(\omega_n, \delta_n).$$

The convergence is uniform since

$$L(\omega_n, \delta_n) \leq h(c^n\,\omega) + g(c^n\,\omega) \leq h(c^n x_0) + g(c^n x_0)$$

and it was assumed that $\Sigma\, h(c^n x_0)$, $\Sigma\, g(c^n x_0) < \infty$. Thus $\Phi(\omega, s)$ is continuous, and assumes a maximum over the compact set of S.

FUNCTIONAL EQUATION FOR THE OPTIMAL STRATEGY

Returning to the general discussion we now deduce the fundamental functional equation.

$$\Phi(\omega, s^*) = \max_s \sum_{n=1}^{\infty} L(\omega_n, \delta_n) \prod_{i=1}^{n-1} P(\delta_1)$$

$$\max_s \left\{ L(\omega, \delta_1) + P(\delta_1) \sum_{n=1}^{\infty} L(\omega_{n+1}, \delta_{n+1}) \prod_{i=1}^{n-1} P(\delta_{i+1}) \right\}$$

$$= \max_s \left\{ L(\omega, \delta_1) + P(\delta_1)\, \Phi(T_{\delta_1}\omega, s') \right\}$$

where $s' = (\delta_2, \delta_3, \delta_4, \dots)$. Let $K(\omega) = \max_s \Phi(\omega, s)$. Then

$$K(\omega) = \max_s \left\{ L(\omega, \delta_1) + P(\delta_1)\, \Phi(t_{\delta_1}\omega, s') \right\}$$

$$= \max_{\delta_1} \max_{s'} \left\{ L(\omega, \delta_1) + P(\delta_1)\, \Phi(T_{\delta_1}\omega, s') \right\} \tag{2}$$

$$= \max_{\delta_1} \left\{ L(\omega, \delta_1) + P(\delta_1)\, K(T_{\delta_1}\omega) \right\}.$$

This is the "Principle of Optimality." The equation is just the general formulation of the functional equation which is used in almost all dynamic programing problems. Specializing it to that of Example 1 for the case of two operators, (2) becomes

$$K(\omega) = f(x, y) = \max \begin{cases} p_1 r_1 x + p_1 f((1 - r_1)x, y) \\ p_2 r_2 y + p_2 f(x, (1 - r_2)y) . \end{cases}$$

For example 2 we obtain

$$K(\omega) = f(x) = \max_{0 \le y \le x} \{ g(y) + h(x - y) + f(ay + b(x - y)) \} . \tag{2a}$$

Uniqueness

We have shown that the optimal return $K(\omega)$ to the problem satisfies the functional equation (2). But actually in most models considered one usually proceeds in the reverse direction. We start with a solution $f(\omega)$ of the functional equation and derive qualitative properties of optimal strategies from it. This requires that we must establish that the solution $f(\omega)$ to the functional equation is $K(\omega) = \max_s \Phi(\omega, s)$. This amounts in effect to establishing the uniqueness of solutions of (2) among a suitable class (containing M) of possible solutions.

To this end, let $M(\omega)$ be a solution to the functional equation

$$M(\omega) = \max_{\delta_1}\left\{ L(\omega, \delta_1) + P(\delta_1)\, M(T_{\delta_1}\, \omega)\right\}$$

$$= \max_{\delta_1}\left\{ L(\omega, \delta_1) + P(\delta_1)\,[\max_{\delta_2}\{L(T_{\delta_1}\, \omega, \delta_2) + P(\delta_2)\, M(\omega_3)\}]\right\}$$

$$= \max_{\delta_1} \max_{\delta_2}\left\{ L(\omega, \delta_1) + P(\delta_1)\, L(T_{\delta_1}\, \omega, \delta_2) + P(\delta_1)\, P(\delta_2)\, M(\omega_3)\right\}.$$

Therefore,

$$M(\omega) = \max_{s} \Phi(\omega, s) + \lim_{n\to\infty} \max_{\delta_1, \ldots, \delta_n} M(\omega_n) \prod_{i=1}^{n-1} P(\delta_i)$$

$$= K(\omega) + \lim_{n\to\infty} \max_{\delta_1, \ldots, \delta_n} M(\omega_n) \prod_{i=1}^{n-1} P(\delta_i). \tag{2}$$

To establish the uniqueness of the solution $M(\omega)$ of (2), we must obtain that the limit term is zero. This need not be true in general for any solution to the functional equation. In each problem we shall restrict ourselves to an appropriate class of M's for which the lim = 0 for any M in the class. In the first example

$$\max_{\delta_1, \ldots, \delta_n} M(\omega_n) \prod_{i=1}^{n-1} P(\delta_i) \le \max_{\delta_1, \ldots, \delta_n} M(\omega_n)\, \alpha^{n-1}.$$

Hence, for all bounded M the lim is 0 as $\alpha < 1$. Thus, in this example when we restrict ourselves to bounded solutions of the functional equation, then the solution is unique. In the second example

$$\max_{\delta_1, \ldots, \delta_n} M(\omega_n) \prod_{i=1}^{n-1} P(\delta_i) \le \max_{\delta_1, \ldots, \delta_n} M(\omega_n)$$

where $\omega_n \to 0$ as $n \to \infty$ so the max $\to 0$ for any M which is continuous and equal to 0 at the origin. This is indeed easily seen to be satisfied for $K(\omega)$ on account of the monotonicity of g and h. Restricting ourselves to the class of M which are continuous and equal to 0 at the origin gives the desired uniqueness.

Remark 1. If instead of requiring uniform convergence of $\sum_n L(\omega_n, \delta_n)\, P_n(s)$ we require that $\Phi(\omega, s)$ be continuous in s, then the convergence must be uniform. This follows from a theorem of Dini, (an increasing sequence of continuous functions converges to a continuous function on a compact space if and only if the convergence is uniform), since each term of the series is positive.

Remark 2. If we further assume that $\Sigma\, L(\omega_n, \delta_n^*)\, P_n(s^*)$ converges uniformly in ω, then $\Phi(\omega, s^*)$ is also a continuous function of ω. The reader can verify for himself that we have uniform convergence in ω in the two examples.

Remark 3. A useful fact is that $M_k(\omega) - M(\omega)$ where M_k is the optimal return if there are just k stages. This follows easily from the uniform convergence of $\sum_n L(\omega_n, \delta_n)\, P_n(s)$. The details shall be omitted.

Simple generalizations of our models can be presented along the following lines: we can allow the decision spaces to vary with each stage as well as the payoff functions. Of course, this corrupts the symmetry of the problem as we pass from one stage to another and the functional equation in (2) is consequently no longer valid. Formally, we can allow the decisions to be made continuously. This possible generalization leads to some mathematical complexities, but in principle it can be handled by our methods. As a recourse we may permit decisions to be made at discrete random times. These possible extensions shall not be pursued further in this paper.

NON-DETERMINISTIC DYNAMIC PROGRAMING PROBLEM

Our next object is to extend all the previous considerations to a nondeterministic phenomenon. The abstract structure takes now the following form:

1. The state space Ω consists of a set of random variables defined on a common sample space with values in a set U. This represents all possible states of nature that can confront us.

2. The decision space D consists of a set of functions mapping U into another space A referred to as the action space. The action space A defines a topological space.

3. The payoff function L is a real valued function $L(u, a)$ defined for each point u in U and a in A.

4. A collection 0 of transition operators T_δ indexed by $\delta \in D$ which map Ω into itself is given.

The same topological assumptions apply as before and a class of functions $P(\delta)$ is also prescribed. The strategy space S is defined as the direct product of a countable number of replications of D. If D is assumed to be compact, then S is compact, as before. The payoff of the strategy s is given by the expression

$$\Phi(\omega, s) = \sum_{n=0}^{\infty} P_n(s)\, \mathcal{E}\,(L(\omega_n, \delta_n(\omega_n))) \tag{3}$$

where $\omega_0 = \omega$ is the initial state of Ω, $\omega_n = T_{\delta_{n-1}}\, \omega_{n-1}$, $\delta_n(\omega_n)$ is the decision defined by the n^{th} component of s, δ_n, and $\mathcal{E}$ is the expected value of the random variable defined by $L(\omega, \delta(\omega))$ with $P_n(s) = \prod_{i=1}^{n-1} P(\delta_i)$. Usually $\omega_0 = \omega$ consists of a random variable with only one possible value (i.e., a known initial state), then the initial term of $\Phi(\omega, s)$ becomes $L(\omega_0, \delta_0(\omega_0))$. Again, the series (3) is assumed to converge uniformly in s for each ω. The pertinent functional equation (principle of optimality) under these circumstances takes the form

$$\max_s \Phi(\omega_0, s) = M(\omega_0) = \max_\delta [L(\omega_0, \delta(\omega_0)) + P(\delta)\, \mathcal{E}\,(M(T_\delta\, \omega_0))]. \tag{4}$$

An analogous argument as before under similar conditions establishes the existence and uniqueness of a solution to the functional equation (4).

Example 3. We apply the above setup to a discussion of existence of optimal strategies for the inventory problem. It is of interest to compare our treatment with that of [4].

Let x, $0 \leq x < \infty$, represent the amount of commodities available at the initial period and suppose one orders y. The demand over a unit interval of time z is a random variable with distribution $F(z)$ which is assumed to be known and independent of the time interval. Let $L_i(x, y) = L(x, y)$ represent the total loss during the i^{th} time interval with x available initially and an ordering policy given by y. We now impose a restriction on the choice of additional ordering y in that we require $0 \leq y \leq M$ where M is a fixed bound. This can be insured for instance without loss of generality by requiring the cost to tend to infinity. It can be justified as realistic. We also suppose that $L(x, y)$ is defined only for $x \geq 0$ and $y \geq 0$. The ordering policy y_i is a random variable depending on the amount of goods x available at the i^{th} state which was a random variable. The space Ω consists of the space of all random variables describing the amount of goods available for any period of time where we start with x $(0 \leq x < \infty)$. Obviously, the set $U = [0 \leq x \leq \infty]$. The action space $A = [0 \leq y \leq M]$ and thus the decision space D consists of all bounded measurable functions $y(x)$ mapping U into A. The space is a compact set viewed in the weak*topology as a subset of the collection of all bounded functions which is in the conjugate space of the Banach space of all integrable functions defined on U. We assume also that $\mathcal{E}(L(x, y(x)))$ is a continuous function of the decision y with respect to the weak*topology imposed on D. This condition is met for instance if the loss is a linear function of the amount ordered and the distribution F is absolutely continuous. Other types of loss functions can be handled with similar techniques by employing other suitable topologies. The transformation $T_y x$ is a mapping of the random variable state at the beginning of the period after initially ordering y into the state of the remainder of the goods at the end of the period. Finally, put $P(\delta) \equiv a$ where $0 < a < 1$ and a is fixed. The total loss for the initial state x following an optimal policy is therefore

$$\mathcal{L}(x) = \inf_{s} [L(x, y_0) + a\,\mathcal{E}[L(t_{y_0} x, y_1)] + a^2\,\mathcal{E}[L(T_{y_1} x_1, y_2)] + \cdots \tag{5}$$

If $\in L(T_y x, \bar{y})$ is uniformly bounded in the decisions y and $\bar{y}$, then (5) converges uniformly and we obtain the existence and uniqueness of the solution to the functional equation (4) where max in that equation is replaced by min. The functional equation can be written in the form

$$\mathcal{L}(x) = \inf_{y \geq x} \{L(x, y) + a\mathcal{L}(0)\,[1 - F(y)] + a \int_0^y \mathcal{L}(y - t)\, dF(t)\}$$

with $\mathcal{L}(x) = M(x)$.

The usual theorems of the convergence of the optimal return for the finite stage problem to the return of the infinite problem remain valid in this case. In a future paper, specific results about the form of the optimal strategy for the inventory model will be presented.

REFERENCES

[1] K. Arrow, T. Harris, and J. Marschak, "Optimal Inventory Policy," Econometrica, Vol. 19 (1951), pp. 250-272

[2] R. Bellman, "Some Problems in the Theory of Dynamic Programing," Econometrica, Vol. 22 (1954), pp. 37-48

[3] R. Bellman, "On the Theory of Dynamic Programing," Proc. Nat. Acad. Sci. Vol. 38 (1952), pp. 716-719

[4] Dvoretzky, A., Kiefer J. and Wolfowitz, J., "The Inventory Problem I: Case of Known Demand," Econometrica, Vol. 20 (1952) pp. 187-222

* * *

NEWS AND MEMORANDA

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics.

RADM Rawson Bennett has been appointed Chief of Naval Research to succeed RADM F. R. Furth, who is retiring. Admiral Bennett was formerly Assistant Chief of the Bureau of Ships for Electronics.

Messrs. John C. Ten Eyck and Frederick C. Dyer are preparing a book titled "Founders of Modern Logistics," and would like to receive suggestions about names to be included and the book's scope and approach. Such suggestions should be sent to Mr. Dyer's home, 4509 Cumberland Avenue, Chevy Chase 15, Maryland.

A joint meeting of the Western Section of the Operations Research Society and the Southern California Chapter of the Institute of Management Science will be held on the campus of the University of California at Los Angeles on March 30 and 31, 1956. Inquiries concerning the meeting should be sent to Dr. Paul Stillson, Ramo-Wooldrige Corporation, 8820 Bellanca Avenue, Los Angeles 45, California.

The primary focus of the group research activities of The Graduate School of Industrial Administration at the Carnegie Institute of Technology, since the School's establishment six years ago, has been the problem of gaining a better understanding of the administrative process, broadly defined. One major segment of this activity, concerned primarily with the improvement of techniques of administrative decision-making, is the project on "Planning and Control of Industrial Operations" carried out under contract with the Office of Naval Research.

The project on "Planning and Control of Industrial Operations" has been carried on by two informal teams, operating on separate sets of problems but with frequent interaction. One team, comprised of Professors W. W. Cooper, A. Charnes, C. Lemke, and the late Professor Alexander Henderson, has directed its main effort toward the development and application of new techniques, particularly those related to linear programming, for handling a variety of problems in industrial planning and control.

The initial work in this area, supported by an Air Force contract for research in intra-firm behavior, was published in the volume An Introduction to Linear Programming (Wiley: 1953). During the past three years the team has continued to extend these tools and their applications in a number of directions. The published reports on this work include papers in the Scientific American, August 1954, Naval Research Logistics Quarterly, December 1954, Management Science, October 1954 and January 1955, and Econometrica, July 1955. The new

types of problems that have been involved in these linear programming applications lie in such diverse areas as plant location, advertising budgets, executive compensation, and warehousing.

The second team investigating the "Planning and Control of Industrial Operations" has been comprised of Professors C. C. Holt, F. Modigliani, H. A. Simon, R. W. Culbertson, and J. F. Muth, together with several graduate assistants: Robert Byrne, Toshiro Makibuchi, Clyde Roberson, and Peter Winters. This team has focussed its effort on the problem of developing mathematical tools for optimal production and inventory decision-making under a range of different circumstances. The research began with a formulation of such problems within the context of a dynamic programming model. The team discovered that if the cost function to be minimized could be approximated by quadratic terms, a linear optimal decision rule could be obtained that was exceedingly simple and convenient from the viewpoints of data requirements, computation, and administrative application. In the past year, the team has been occupied with a test application of its procedures in an actual operating situation, and with the extension of the quadratic cost criterion techniques to a broader class of problems. Some preliminary reports on this work have appeared in the Journal of the Operations Research Society of America, August 1954, and ONR Research Reviews, December 1954. A more complete description of the procedures will appear in Management Science, October 1955 and January 1956, and Econometrica, April 1956.

Work closely paralleling the studies described above has been carried on in the Graduate School of Industrial Administration under research grants from two business firms—Westinghouse Airbrake Company and the public accounting firm of Touche, Niven, Bailey, and Smart. The first of these is an exploratory research program in the transportation industry, which will be especially concerned with the development and application of mathematical and statistical tools and approaches to problems of the railroads. Faculty members associated with this project include Professors H. Wein, M. Miller, A. Charnes, and E. Mansfield. The second project has been an investigation of the application of statistical techniques to problems of management controls, notably in the accounting and auditing areas. Professors W. W. Cooper and R. M. Cyert have worked in this area with staff members of Touche, Niven, Bailey, and Smart. Some of the findings of the study have been published in the Accounting Review, March 1954 and April 1955, and Cost and Management, May 1955.

All of the projects described above approach logistics from the standpoint of the mathematician and econometrician, and aim primarily at the construction of normative rules to guide the decision maker. A second group of projects in the Graduate School of Industrial Administration has taken as its starting point the administrative and organizational problems involved in decision making, and has sought to attack them with the assistance of tools drawn from social psychology and sociology. Four principal projects have been completed or are currently under way in this area:

1. A study of the ways in which the organization of the company's accounting department influences the use of accounting information by operating executives has been reported in the volume, Centralization vs. Decentralization in Organizing the Controller's Department. The project staff included Professors H. Guetzkow, G. Kozmetsky, H. A. Simon, and D. G. Tyndall, and the study was supported by a research grant from the Controllership Foundation.

2. An Organization Behavior Laboratory, under the direction of Professor H. Guetzkow, studies artificial "organization-like" groups under laboratory conditions—thus permitting the test, with adequate controls, of hypotheses derived from organizations in the field. A report on one of the laboratory studies is published in Management Science, April-July 1955.

3. An inventory is being made, by Professors H. Guetzkow, J. March, and H. A. Simon, of the current status of theory and evidence relating to human behavior in organizations.

4. A large-scale study is under way aimed at obtaining a deeper understanding of "judgmental" decision processes at high executive levels in organizations. Arrangements have been made to maintain an observer in each of four industrial concerns. He will follow in detail the steps involved in making one or more policy decisions. The staff of the study includes Professors H. A. Simon, R. M. Cyert, W. Dill, C. J. Haberstroh, and D. B. Trow. This study and the previous one are financed by research grants from the Ford Foundation.

The Managing Editor wishes to thank the following individuals who assisted the members of the Editorial Staff in the refereeing of papers for volume 2: K. Arrow, I. Glicksberg, J. Harrison, U. Henderson, W. Jacobs, R. McShane, R. Norman, W. Orchard-Hays, M. Richardson, M. Rose, J. Sacks, M. Shubik, H. Solomon, P. Wolfe, and J. Wolfowitz.

* * *

RECENT PUBLICATIONS

NORTH AMERICAN SUPPLY. By H. Duncan Hall, Her Majesty's Stationery Office and Longman, Green and Co., London, 1954, 559 pp.

This book describes the history of the contribution of North American Supply to British War Production. While, in terms of monetary value, the contribution of the United States and Canada was smaller than the public generally imagined it to be (seventeen percent and ten percent respectively) its importance was certainly crucial, particularly in machine tools and merchant ships. The book was written principally in Washington to take advantage of the presence there of the bulk of the relevant records, especially the archives of the British Supply Council in North America, which functioned essentially as coordinating body of the various British Missions. It traces the logistic history of the collaboration of the partners, from neutrality through lend-lease to the swift termination of the machinery of the British-America combination with the end of the war, with principal emphasis on the economic and political aspects of supply problems.

UNCERTAINTY AND BUSINESS DECISIONS. Edited by C. F. Carter, G. P. Meredith and G. L. S. Shackle, Liverpool: University of Liverpool Press, 1954, viii + 104 pp.

This book is a symposium on Professor G. L. S. Shackle's theory of expectation which was first propounded in Expectation in Economics in 1949. The contributors are specialists in the fields of economics, mathematics, philosophy, and psychology. The book contains the following papers: I. Uncertainty as a Philosophical Problem (I) - Professor W. B. Gallie; II. Uncertainty as a Philosophical Problem (II) - Dr. D. J. O'Connor; III. The Appropriate Mathematical Tools for Describing and Analyzing Uncertainty - Dr. I. J. Good; IV. Methodological Considerations in the Study of Human Anticipation - Professor G. P. Meredith; V. A Revised Theory of Expectations - Professor C. F. Carter; VI. The Impact of Uncertainty on Economic Theory, with Special Reference to the Theory of Profits - Professor B. R. Williams; VII. The Possibility of Empirical Testing of Alternative Theories of Uncertainty - A. D. Roy; VIII. Uncertainty, Expectation and Decision in Business Affairs - Professor G. L. S. Shackle; IX. Expectations in Economics - Professor G. L. S. Shackle. In addition, there is a note on an empirical study of business decision-making made in the early thirties by Mr. W. R. Dunlop.

MATHEMATICAL THINKING IN THE SOCIAL SCIENCES. Edited by Paul F. Lazarsfeld, Glencoe, Illinois: The Free Press, 1954, 444 pp.

The purpose of this volume is to present a sampling of situations in the domain of the social scientist which are profitably explored with the aid of mathematics. Each of the eight papers, apart from exploring specific problems, attempts to offer some insight into the role mathematics can play in the social sciences. The audience for which the volume is mainly

slanted is the social scientist with a modicum of training in mathematics. The authors of the various papers are T. W. Anderson, Nicholas Rashevsky, James S. Coleman, Jacob Marschak, Louis Guttman, and Paul F. Lazarsfeld. The Introduction, which includes a summary of each of the papers, was written by the editor.

THE UNITED STATES AND WORLD SEA POWER. Edited by E. B. Potter and J. R. Fredland, Prentice Hall, Inc., Englewood Cliffs, N. J., 1955, 963 pp.

This work, prepared by the naval history faculty of the United States Naval Academy, is a history of United States sea power, developed as part of the history of international sea power and in the context of world history. In the editors' words, the book "develops six main themes: (1) the influence of sea power upon history, (2) the rationale of strategic decision, (3) the characteristics of successful leadership, (4) the development of naval weapons, (5) the evolution of naval tactics, and (6) the evolution of amphibious doctrine. Emphasis is placed on the problems posed in each period of history by new weapons and new conditions, and on the solutions worked out for each by the navies of the world. The reader will note that problems of a similar or parallel nature arise in each epoch and that these have to be solved in most instances by somewhat similar means. He will thus perceive the parallels of naval history and the value of studying the past as a guide to future action. At the same time he will note that the problems are never precisely the same and will perceive thereby the dangers inherent in trying to fight the last war."

UNITED STATES ARMY IN WORLD WAR II - The Technical Services - The Ordnance Department: Planning Munitions for War. By Constance McLaughlin Green, Harry C. Thomson and Peter C. Roots, Office of the Chief of Military History, Department of the Army, Washington, D. C., 1955, 542 pp.

This is the first volume of a projected three volume history of the Ordnance Corps in World War II, and treats the steps that precede munitions manufacture. The book begins with the background of pre-1940 difficulties and the period when the United States was neither at peace nor at war. The second section describes the evolution of a workable organization and the recruiting and training of soldiers and civilians to carry out the Ordnance mission. The last section deals with research and development of weapons, ammunition, armor and vehicles.

* * *

☆ U. S. GOVERNMENT PRINTING OFFICE : 1956 O— 379857

22035 15